Mathematical Population Genetics

Introduction to the Retrospective View of Population Genetics Theory

Lecture Notes

Math 563

Paul Joyce

Introduction

We begin by introducing a certain amount of terminology from population genetics.

Every organism is initially, at the time of conception, just a single cell. It is this cell, called a *zygote* (and others formed subsequently that have the same genetic makeup), that contain all the relevant genetic information about an individual and influences that of its offspring. Thus, when discussing the genetic composition of a population, it is understood that by the genetic properties of an individual member of the population one simply means the genetic properties of the zygote from which the individual developed.

Within each cell are a certain fixed number of *chromosomes*, threadlike objects that govern the inheritable characteristics of an organism. Arranged in linear order at certain positions, or *loci*, on the chromosomes, are *genes*, the fundamental units of heredity. At each locus there are several alternative types of genes that can occur; the various alternatives are called *alleles*.

Diploid organisms are those for which the chromosomes occur in *homologous* pairs, two chromosomes being homologous if they have the same locus structure. An individual's genetic makeup with respect to a particular locus, as indicated by the unordered pair of alleles situated there, (one on each chromosome), is referred to as its *genotype*. Thus if there are 2 alleles A, B at a given locus, there are 3 genotypes AA, AB, BB.

Haploid organisms, are those for which there is a single chromosome in each cell. While most organisms of interest are not haploid, if a gene is only maternally or paternally inherited, then for the purposes of studying that locus one can treat the population as haploid.

The Effects of Genetic Drift

We will assume a reproductive process called the *Wright-Fisher model*. The reproductive process can be roughly described as follows.

1. The generations are non overlapping.
2. The population size N is fixed throughout.
3. Each individual has a large number of *gametes*, haploid cells of the same gene (neglecting mutation) as that of the zygote. We suppose that the number of gametes is (effectively) infinite, and are produced without fertility differences, that is, that all genotypes have equal probabilities of transmitting gametes in this way. The next generation is produced by sampling N individuals.

The following discussion uses concepts from probability. If you need to review these concepts then you might want to visit the WEB site below. The site contains an online statistics text. The text is written by Philip B. Stark Department of Statistics University of California, Berkeley. Material from Chapters 9, 11 12 are used in the discussion that follows.

An online statistics text.
http://www.stat.Berkeley.EDU/users/stark/SticiGui/Text/index.htm

Random genetic drift: The mechanism by which genetic variability is lost through the effects of random sampling.

Objective Questions

1. How likely is an allele to be lost due to genetic drift?
2. How long on average does it take an allele to fix in the population?

For now we restrict attention to the 2 allele Wright-Fisher model with no mutation. Denote the alleles by A, B. Let X_n be the number of A alleles in the population in generation n. Assume a population size is N. Denote by

$$p_{ij} = P(X_n = j | X_{n-1} = i)$$

The Wright Fisher model assumptions are equivalent to having each individual in generation n choose a parent from generation $n-1$ at random with replacement. That is, if there are i alleles of type A in generation $n-1$ then the probability π_i that a particular individual will be of type A in generation n is $\pi_i = i/N$. Therefore, p_{ij} follows a binomial probability distribution given by

$$p_{ij} = \binom{N}{j} (\pi_i)^j (1-\pi_i)^{N-j}, \quad 0 \le i, j \le N. \tag{1}$$

Exercise 1.1 Suppose $N = 3$, Use Equation (1) to calculate p_{ij} for each of the 16 possibilities. Write your answer in a 4×4 matrix. Calculate $P(X_2 = 0 | X_0 = 2)$.

From the nature of the binomial distribution we have

$$E(X_n | X_{n-1} = i) = N \frac{i}{N} = i. \tag{2}$$

By the definition of conditional expectation we have

$$E(X_n | X_{n-1} = i) = \sum_{j=1}^{N} j P(X_n = j | X_{n-1} = i) = \sum_{j=1}^{N} j p_{ij} \tag{3}$$

By setting the right hand side of Equation (2) equal to the right hand side of (3) we get

$$i = \sum_{j=1}^{N} j p_{ij} \tag{4}$$

It follows from (2) that

$$\begin{aligned} E(X_n) = & \sum_{i=1}^{N} E(X_n|X_{n-1}=i)P(X_{n-1}=i) \\ & = \sum_{i=1}^{N} iP(X_{n-1}=i) \\ & = E(X_{n-1}). \end{aligned}$$

By the same argument $E(X_{n-1}) = E(X_{n-2})$, so we can conclude

$$E(X_n) = E(X_{n-1}) = \cdots = E(X_0).$$

This result can be thought of as the analog of the Hardy-Weinberg observation that in an infinitely large random mating population, the relative frequencies of the alleles remains constant in every generation.

We are now ready to answer question (1). We begin by calculating a_i which is the probability that eventually the population contains only A alleles, given that $X_0 = i$. The standard way to find such a probability is to condition on the value of X_1. That is, the population could go from i type A alleles to j type A alleles in the first generation and then the j type A alleles eventually become fixed. Considering each j between 0 and N gives

$$a_i = \sum_{j=0}^{N} p_{ij} a_j \tag{5}$$

where $a_0 = 0$ and $a_N = 1$. Now recall equation (4). Divide both sides of (4) by N to get

$$i/N = \sum_{j=0}^{N} p_{ij}(j/N).$$

Therefore $a_i = i/N$ is a solution to (5). Since (5) represents a system of $N-1$ linear equations with $N-1$ unknowns, it is not difficult to show that the above solution is unique. We have now shown that the probability that a particular allele with i representatives will eventually fix in a population of size N is i/N. Therefore, the probability that an allele with frequency i will be lost due to genetic drift is $1 - i/N$.

We have seen that variability is lost from the population. How long does fixation take? First we find an equation satisfied by m_i, the mean time to fixation starting from $X_0 = i$. To do this, notice first that $m_0 = m_N = 0$. Now condition on the first step. If the population goes from i type A alleles to j A (assume $j \neq N$) alleles in the first generation, then the mean time to fixation is $1 + m_j$. Averaging over all possible first generations gives

$$m_i = p_{i0} \cdot 1 + p_{iN} \cdot 1 + \sum_{j=1}^{N-1} p_{ij}(1 + m_i) = 1 + \sum_{j=0}^{N} p_{ij} m_j \tag{6}$$

Exercise 1.2 Solve Equation (6) when $N = 3$.

For even moderate size N Equation (6) becomes very complicated. An approximation, valid for large N is the following

$$m_i \approx -2(i \log(i/N) + (N - i) \log(1 - i/N)).$$

If you are interested in the details for deriving the above equation, see me. If $i/N = 1/2$ then

$$m_i \approx -2N \log(1/2) = 2N \log 2 \approx 1.39N$$

whereas

$$m_1 \approx 2 \log(N).$$

The effects of Mutation

Objective Question

1. How does the process of mutation maintain genetic variability?

We suppose that a probability $\mu_A > 0$ that an A allele mutates to a B allele in a single generation, and the probability $\mu_B > 0$ that a B allele mutates to an A. The stochastic model for X_n given X_{n-1} is the same described in (1), but where

$$\pi_i = \frac{i}{N}(1 - \mu_A) + \left(1 - \frac{i}{N}\right) \mu_B.$$

Mutation changes the nature of the process in a very significant way. In the no mutation model, eventually one allele fixes in the population and the other allele goes away. In mathematics we refer to this as an absorbing boundary. In the process with mutation from A to B and B to A one allele can never remain fixed forever. If by chance the population is, at some point in time, fixed with type A, then eventually one of the alleles in the population will mutate back to a B. Processes of this type have the property that as $n \to \infty$ then X_n converges to a random variable X. We call the probability distribution of X the stationary distribution.

We will now calculate the $E(X)$. First note that by the same argument used in (2) we get

$$E(X_n|X_{n-1} = i) = N\pi_i = N\left(\frac{i}{N}(1 - \mu_A) + \left(1 - \frac{i}{N}\right)\mu_B\right) = i(1-\mu_A-\mu_B)+N\mu_B.$$

Therefore,

$$E(X_n|X_{n-1}) = X_{n-1}(1 - \mu_A - \mu_B) + N\mu_B.$$

and

$$E(X_n) = E(E(X_n|X_{n-1})) = E(X_{n-1})(1 - \mu_A - \mu_B) + N\mu_B$$

At stationarity $\lim_{n\to\infty} E(X_n) = \lim_{n\to\infty} E(X_{n-1}) = E(X)$. This implies

$$E(X) = E(X)(1 - \mu_A - \mu_B) + N\mu_B$$

and solving for $E(X)$ gives

$$E(X) = \frac{N\mu_B}{\mu_B + \mu_A}.$$

Prospective versus Retrospective

The theory of population genetics developed in the early years of this century focused on a *prospective* treatment of genetic variation. Given a stochastic model for the evolution of gene frequencies one can ask questions like 'How long does a new mutant survive in the population?' 'What is the chance that an allele becomes fixed in the population?'. These questions involve the analysis of the future behavior of a system given initial data. In this section we studied the two allele Wright-Fisher model. As a result we got a taste of the prospective treatment for this simple model. Most of the theory is much easier to think about if the focus is *retrospective*. Rather than ask where the population will go, ask where it has been. We shall see that the retrospective approach is very powerful and technically simpler. In the rest of the notes we will take this view.

The coalescent

Introduction

In 1982 John Kingman, inspired by his friend Warren Ewens, took to heart the advice of Danish philosopher Soren Kierkegaard and realized that "Life can only be understood backwards, but it must be lived forwards." Applying this perspective to the world of population genetics led him to the development of the *coalescent*, a mathematical model for the evolution of a sample of individuals drawn from a larger population. The coalescent has come to play a fundamental role in our understanding of population genetics and has been at the heart of a variety of widely-employed analysis methods. For this it also owes a large debt to Richard Hudson, who arguably wrote the first paper about the coalescent that the non-specialist could easily understand. Here we introduce the coalescent, summarize its implications, and survey its applications.

The central intuition of the coalescent is driven by parallels with pedigree-based designs. In those studies, the shared ancestries of the sample members, as described by the pedigree, are used to inform any subsequent analysis, thereby increasing the power of that analysis. The coalescent takes this a step further by making the observation that there is no such thing as unrelated individuals. We are all related to some degree or other. In a pedigree the relationship is made explicit. In a population-based study the relationships are still present, albeit more distant, but the details of the pedigree are unknown. However, it remains

the case that analyses of such data are likely to benefit from the presence of a model that describes those relationships. The coalescent *is* that model.

Motivating problem

Human evolution and the infinitely- many-sites model

One of the signature early applications of the coalescent was to inference regarding the early history of humans. Several of the earliest data-sets consisted of short regions of mitochondrial DNA [mtDNA] or Y chromosome. Since mtDNA is maternally inherited it is perfectly described by the original version of the coalescent, with its reliance upon the existence of a single parent for each individual and its recombination-free nature. To motivate what follows, here we consider one of those early data sets.

The data in the following example comes from Ward *et. al.* (1991). The data analysis and mathematical modeling comes from a paper by Griffiths and Tavaré (1994).

Mitochondria DNA (mtDNA) comprises only about 0.00006% of the total human genome, but the contribution of mtDNA to our understanding of human evolution far outweighs its minuscule contribution to our genome. Human mitochondrial DNA, first sequenced by Anderson *et.al.* (1981), is a circular double-stranded molecule about 16,500 base pairs in length, containing genes that code for 13 proteins, 22 tRNA genes and 2 rRNA genes. Mitochondria live outside the nucleus of cells. One part of the molecule, the control region (sometimes referred to as the D-loop), has received particular attention. The region is about 1,100 base pairs in length.

As the mitochondrial molecule evolves, mutations result in the substitution of one of the bases A,C,G or T in the DNA sequence by another one. Transversions, those changes between purines (A,G) and pyrimidines (C,T), are less frequent than transitions, the changes that occur between purines or between pyrimidines.

It is known that base substitutions accumulate extremely rapidly in mitochondrial DNA, occurring at about 10 times the rate of substitutions in nuclear genes. The control region has an even higher rate, perhaps on order of magnitude higher again. This high mutation rate makes the control region a useful molecule with which to study DNA variation over relatively short time spans, because sequence differences will be found among closely related individuals. In addition, mammalian mitochondria are almost exclusively maternally inherited, which makes these molecules ideal for studying the maternal lineages in which they arise. This simple mode of inheritance means that recombination is essentially absent, making inferences about molecular history somewhat simpler than in the case of nuclear genes.

In this example, we focus on mitochondrial data sampled from a single North American Indian tribe, the Nuu-Chah-Nulth from Vancouver Island. Based on the archaeological records (cf. Dewhirst, 1978), it is clear that there is a

remarkable cultural continuity from earliest levels of occupation to the latest. This implies not only that there was no significant immigration into the area by other groups, but that subsistence pattern and presumably the demographic size of the population has also remained roughly constant for at least 8,000 years. Based on the current size of the population that was sampled, there are approximately 600 women of child bearing age in the traditional Nuu-Chah-Nulth population.

The original data, appearing in Ward *et. al.* (1991) comprised a sample of mt DNA sequences from 63 individuals. The sample approximated a random sample of individuals in the tribe, to the extent to which this can be experimentally arranged. Each sequence is the first 360 basepair segment of the control region. The region comprises 201 pyrimidine sites and 159 purine sites; 21 of the pyrimidine sites are variable (or segregating), that is, not identical in all 63 sequences in the sample. In contrast, only if 5 of the purine sites are variable. There are 28 distinct DNA sequences (hereafter called lineages) in the data. Because, no transversions are seen in these data each DNA site is binary, having just two possible bases at each site.

To keep the presentation simple, we focus on one part of the data that seems to have a relatively simple mutation structure. We shall assume that substitutions at any nucleotide position can occur only once in the ancestry of the molecule. This is called **the infinitely- many-sites assumption.** Hence we have eliminated lineages in which substitutions are observed to have occurred more than once. The resulting subsample comprises 55 of the original 63 sequences, and 352 of the original 360 sites. Eight of the pyrimidine segregating sites were removed resulting in a set of 18 segregating sites in all; 13 of these sites are pyrimidines, and 5 are purines. These data are given in Table 1, subdivided into sites containing purines and pyrimidines. Each row of the table represents a distinct DNA sequence, and the frequency of these lineages are given in the right most column of the table.

What structure do these sites have? Because of the infinitely-many-sites assumption, the pattern of segregating sites tells us something about the mutations that have occurred in the history of the sample. Next we consider an ancestral process that could have given rise the observed pattern of variability. This is called the coalescent.

The coalescent process, which we will discuss in some detail as the course goes on, is a way to describe the ancestry of the sample. The coalescent has a very simple structure. Ancestral lines going backward in time coalesce when they have a common ancestor. Coalescence occur only between pairs of individuals. This process may also be thought of as generating a binary tree, with the leaves representing the sample sequences and the vertices where ancestral lines coalesce. The root of the tree is the most recent common ancestor (MRCA) of the sample.

Example 1 To get an idea for building trees from sequences we begin with a simple example. Consider just the segregating purines of lineage b, c, d, e from

Table 1: Nucleotide positions in control region. (Mithocondrial data from Ward et al. (1991, figure1). Variable purine and prymidine positions in the control region. Position 69 corresponds to position 16,092 in the human reference sequence published by Anderson et al. (1981).)

	1	1	2	2	3			1	1	1	1	1	2	2	2	2	3	3	
Position	0	9	5	9	4	8	9	2	4	6	6	9	3	6	7	7	1	3	
	6	0	1	6	4	8	1	4	9	2	6	4	3	7	1	5	9	9	
Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Lineage Frequencies
Lineage																			
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	6
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

Table 1. Below is this reduced data set.

	Site	1	2	3	4
lineage					
b		A	G	G	A
c		G	A	G	G
d		G	G	A	G
e		G	G	G	A

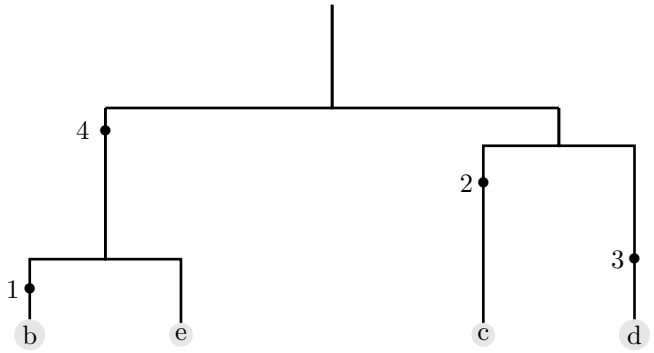

Figure 1: A tree consistent with the data in Example 1.

Suppose $G\ G\ G\ G$ is the ancestral sequence to the four lineages. Figure 1 represents one possible evolutionary scenario connecting the individuals in the sample. The $\bullet$ represents a mutation. The number next to the $\bullet$ represents the position of the mutation. We read the tree diagram in Figure 1 as follows. Start at the ancestral tip of the tree. The first event to occur is a split in the ancestral line. Next a mutation occurs and a G mutates to an A at position 4. This mutation is passed on to lineage b and e. The tree splits again and a mutation occurs at site 2 in lineage c. Next, a mutation occurs at site 3 in lineage d. The final split in the tree separates lineage b and e. The last evolutionary event is a mutation at site 1 in lineage b.

A coalescent tree consistent with the data in Table 1 is given below

A rooted and unrooted gene tree consistent with the data in Table 1 is given below.

Exercise 1 Convince yourself that Figure 2. represents a coalescent tree that is consistent with the data given in Table 1. Use the most frequently occurring

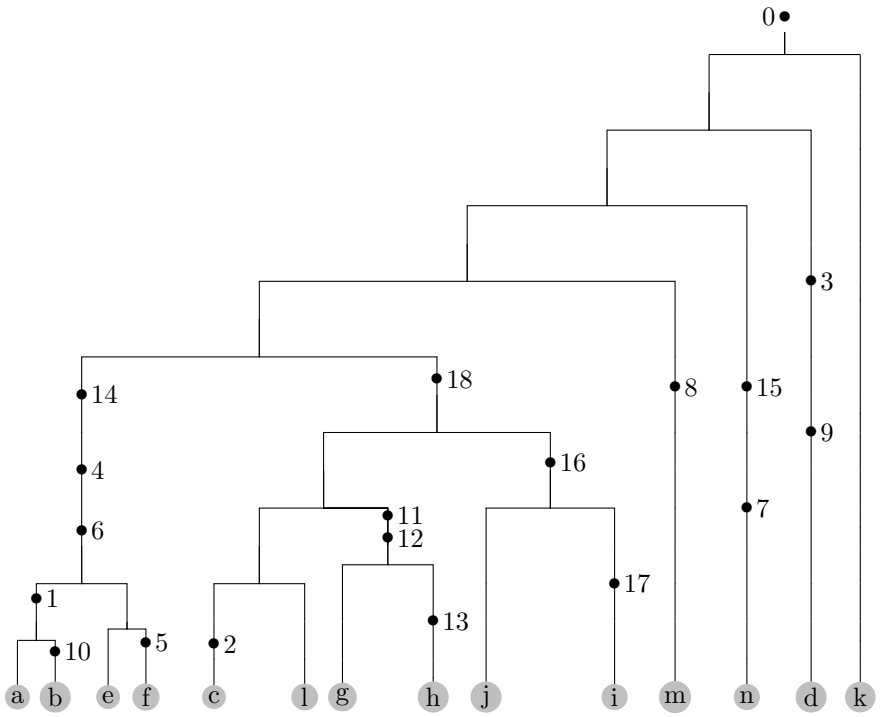

Figure 2: A tree consistent with data in Table 1.

basepair at each site as the ancestral sequence. Describe each event that lead to the sample. Construct another coalescent tree that is consistent with the data.

Exercise 2 Since mutations can occur only once in a given site, there is an ancestral type and a mutant type at each segregating site. For the moment assume we know which is which, and label the ancestral type is 0 and the mutant type as 1. To fix ideas, take each column of the data in Table 1 and label the most commonly occurring base as 0, the other as 1. Construct a matrix of 0's and 1's for the data in Table 1 in the manner described above.

The matrix of 0's and 1's can be represented by a rooted tree by labeling each distinct row by a sequence of mutations up to the common ancestor. These mutations are the vertices in the tree. This rooted tree is a condensed description of the coalescent tree with its mutations, and it has no time scale in it. Figure 3 represents a rooted condensed tree consistent with the data in table 1.

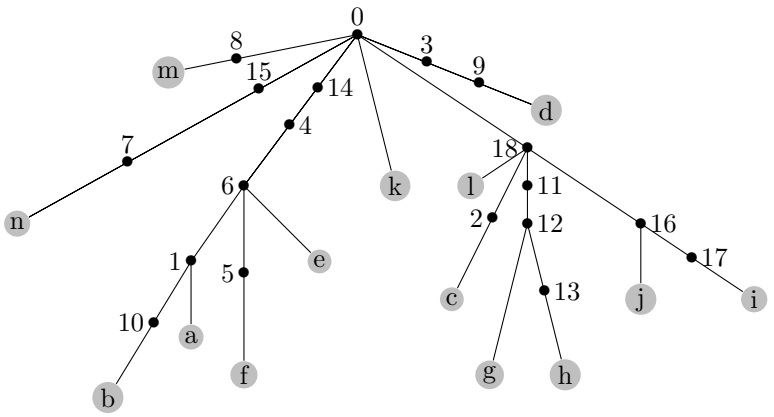

Figure 3: Rooted gene tree.

Exercise 3 Verify that Figure 3 and Figure 4 are consistent with the data. Take lineages a, b, c, d, j and draw the rooted condensed tree for this subset of individuals.

Of course, in practice we never know which type at a site is ancestral. All that can be deduced then from the data are the number of segregating sites between each pair of sequences. In this case the data is equivalent to an *unrooted* tree whose vertices represent distinct lineages and whose edges are labeled by mutations between lineages. The unrooted tree corresponding to the rooted tree in Figure 3 is shown in Figure 4. All possible rooted trees may be found from an unrooted tree by placing the root at a vertex or between mutations, then reading off mutation paths between lineages and the root.

Exercise 4 Construct three rooted trees consistent with the unrooted tree in Figure 4.

We will return to this example later in the course. At that time we will address the problem of estimating the mutation rate, and predicting the time back to the most recent common ancestor. The example illustrates how to connect DNA sequence data to the ancestry of the individuals in the population.

Two technical results

It is convenient to start with two technical results, one of which will be relevant for approximations in the coalescent associated with the Wright-Fisher model.

We consider first a Poisson process in which events occur independently and randomly in time, with the probability of an event in $(t, t + \delta t)$ being $a\delta t$.

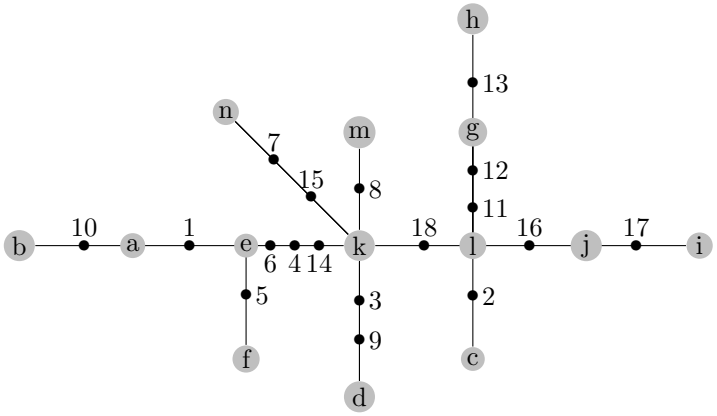

Figure 4: Unrooted gene tree.

(Here and throughout we ignore terms of order $(\delta t)^2$.) We call a the rate of the process. Standard Poisson process theory shows that the density function of the (random) time X between events, and until the first event, is $f(x) = a\ e^{-ax}$, and thus that the mean time until the first event, and also between events, is $1/a$.

Consider now two such processes, process (a) and process (b), with respective rates a and b. From standard Poisson process theory, given that an event occurs, the probability that it arises in process (a) is $a/(a+b)$. The mean number of "process (a)" events to occur before the first "process (b)" event occurs is a/b. More generally, the probability that j "process (a)" events occur before the first "process (b)" event occurs is

$$\frac{b}{a+b}\Big(\frac{a}{a+b}\Big)^j, \quad j = 0, 1, \ldots. \tag{7}$$

The mean time for the first event to occur under one or the other process is $1/(a+b)$. Given that this first event occurs in process (a), the conditional mean time until this first event occurs is equal to the unconditional mean time, namely $1/(a+b)$. The same conclusion applies if the first event occurs in process (b).

Similar properties hold for the geometric distribution. Consider a sequence of independent trials and two events, event A and event B. The probability that one of the events A and B occurs at any trial is $a+b$. The events A and B cannot both occur at the same trial, and given that one of these events occurs at trial i, the probability that it is an A event is $a/(a+b)$.

Consider now the random number of trials until the first event occurs. This random variable has geometric distribution, and takes the value i, $i = 1, 2, \ldots$, with probability $(1-a-b)^{i-1}(a+b)$. The mean of this random variable is thus $1/(a+b)$. The probability that the first event to occur is an A event is $a/(a+b)$. Given that the first event to occur is an A event, the mean number of trials

before the event occurs is $1/(a+b)$. In other words, this mean number of trials applies whichever event occurs first. The similarity of properties between the Poisson process and the geometric distribution is evident.

The coalescent model - no mutation

With the above results in hand, we first describe the general concept of the coalescent process. To do this, we consider the ancestry of a sample of n genes taken at the present time. Since our interest is in the ancestry of these genes, we consider a process moving backward in time, and introduce a notation acknowledging this. We consistently use the notation τ for a time in the past before the sample was taken, so that if $\tau_2 > \tau_1$, then τ_2 is further back in the past than is τ_1.

We describe the common ancestry of the sample of n genes at any time τ through the concept of an equivalence class. Two genes in the sample of n are in the same equivalence class at time τ if they have a common ancestor at this time. Equivalence classes are denoted by parentheses: Thus if $n = 8$ and at time τ genes 1 and 2 have one common ancestor, genes 4 and 5 a second, and genes 6 and 7 a third, and none of the three common ancestors are identical, the equivalence classes at time time τ are

$$(1,2), \quad (3), \quad (4,5), \quad (6,7), \quad (8). \tag{8}$$

Such a time τ is shown in Figure 5.

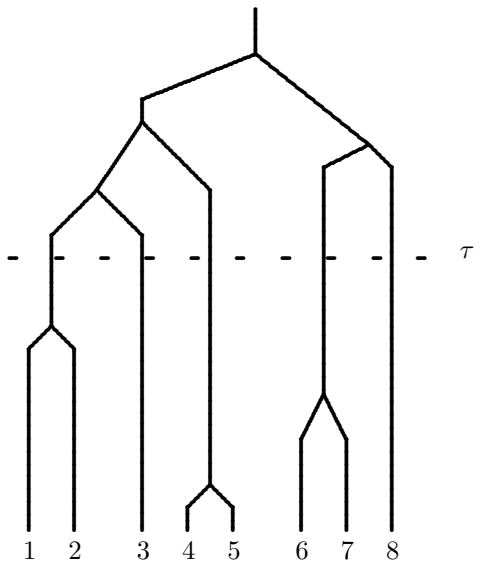

Figure 5: The coalescent

We call any such set of equivalence classes an equivalence relation, and denote any such equivalence relation by a Greek letter. As two particular cases, at time $\tau = 0$ the equivalence relation is $\phi_1 = \{(1), (2), (3), (4), (5), (6), (7), (8)\}$, and at the time of the most recent common ancestor of all eight genes, the equivalence relation is $\phi_n = \{(1, 2, 3, 4, 5, 6, 7, 8)\}$. The coalescent process is a description of the details of the ancestry of the n genes moving from ϕ_1 to ϕ_n.

Let ξ be some equivalence relation, and η some equivalence relations that can be found from ξ by amalgamating two of the equivalence classes in ξ. Such an amalgamation is called a coalescence, and the process of successive such amalgamations is called the coalescence process. It is assumed that, if terms of order $(\delta\tau)^2$ are ignored, and given that the process is in ξ at time τ,

$$\text{Prob (process in } \eta \text{ at time } \tau + \delta\tau) = \delta\tau, \tag{9}$$

and if j is the number of equivalence classes in ξ,

$$\text{Prob (process in } \xi \text{ at time } \tau + \delta\tau) = 1 - \frac{j(j-1)}{2}\delta\tau. \tag{10}$$

The above assumptions are clearly approximations for any discrete-time process, but they are precisely the assumptions needed for the Wright-Fisher approximate coalescent theory. However, the rates are determined by considering a time scale. We know investigate this time scale.

Time Scale

In an evolutionary setting the most convenient way to think about time is to count the number of generations. To convert from generations to years we simply multiply be the mean lifespan of the species. For instance, if two individuals have a common ancestor 10 generations into the past, and the average lifespan of the species is 30 years, then the common ancestor lived (roughly) 300 years ago.

A mathematically convenient time scale is to measure time in units of N generations, where N is the population size. In this time scale one unit of time is equal to N generations.

Example 3.1 In a population of size 1000, if $t = 1/2$ a unit of time has elapsed then 500 generations have passed.

Note that in this time scale a small value for t can represent a fairly large number of generations. We will typically derive results under the mathematically convenient time scale and then convert back to years or generations.

Ancestry of a Sample

Reproduction The most celebrated model for reproduction in population genetics is the Wright-Fisher model. Recall the assumptions of the model are as

follows.

Each individual in the present generation chooses its parent at random from the N individuals of the previous generation. Choices for different individuals are independent, and independent across generations.

Individuals are related through their ancestry and ancestry is revealed through observed mutations. It is convenient to initially separate the mutation process from the ancestral process. For now we will consider only the ancestral process associated with the Wright-Fisher model of reproduction.

First consider a sample of size $n = 2$. How long does it take for the sample of two genes to have its first common ancestor? First we calculate

$$P(2 \text{ individuals have } 2 \text{ distinct parents }) = \left(1 - \frac{1}{N}\right).$$

Since those parents are themselves a random sample from their generation, we may iterate this argument to see that

$$P(\text{First common ancestor more than } r \text{ generations ago}) = \left(1 - \frac{1}{N}\right)^r$$

Let T_2 be the number of generations back into the past until two individuals have a common ancestor, then

$$P(T_2 > r) = P(\text{First common ancestor more than } r \text{ generations ago}) = \left(1 - \frac{1}{N}\right)^r$$

By rescaling time in units of N generations, so that $r = Nt$,and letting $N \to \infty$ we see that this probability is

$$\left(1 - \frac{1}{N}\right)^{Nt} \to e^{-t}$$

Thus the time until the first common ancestor of the sample of two genes has approximately an exponential distribution with mean 1. what can be said of a sample of three genes? We see that the probability that the sample of three has distinct parents is

$$\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)$$

Let T_3 be the number of generations back in time until a two of three genes have a common ancestor. Applying the iterative argument above one more time, we see that

$$\begin{aligned} P(T_3 > r) &= \left[\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\right]^r \\ &= \left(1 - \frac{3}{N} + \frac{2}{N^2}\right)^r \end{aligned}$$

Rescaling time once more in units of N generations, and taking $r = Nt$, shows that for large N this probability is approximately e^{-3t}.

Now consider a sample of size n drawn from a population of size N evolving according the assumptions of the Wright-Fisher Model. The stochastic process that describes the ancestry of the sample is called the n *coalescent*. It is a mathematical approximation, valid for large N.

The coalescent process defined by (9) and (10)consists of a sequence of $n-1$ Poisson processes, with respective rates $j(j-1)/2,\ j = n, n-1, \ldots, 2$, describing the Poisson process rate at which two of these classes amalgamate when there are j equivalence classes in the coalescent. Thus the rate $j(j-1)/2$ applies when there are j ancestors of the genes in the sample for $j < n$, with the rate $n(n-1)/2$ applying for the actual sample itself.

The Poisson process theory outlined above shows that the time T_j to move from an ancestry consisting of j genes to one consisting of $j-1$ genes has an exponential distribution with mean $2/\{j(j-1)\}$. Since the total time required to go back from the contemporary sample of genes to their most recent common ancestor is the sum of the times required to go from j to $j-1$ ancestor genes, $j = 2, 3, \ldots, n$, the mean $E(T_{\text{MRCAS}})$ is, immediately,

$$T_{\text{MRCAS}} = T_n + T_{n-1} + \cdots + T_2 \tag{11}$$

It follows that

$$\begin{aligned} E(T_{\text{MRCAS}}) &= \sum_{k=2}^{n} E(T_k) \\ &= \sum_{k=2}^{n} \frac{2}{k(k-1)} \\ &= 2\sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) \\ &= 2\left(1 - \frac{1}{n}\right) \end{aligned} \tag{12}$$

Therefore

$$1 = E(T_2) \leq E(T_{\text{MRCAS}}) < 2$$

Note that T_{MRCAS} is close to 2 even for moderate n.

Example 2 Again consider a sample of $n = 30$ Nuu-Chah females in a population of $N = 600$. The mean time to a common ancestor of the sample is $2(1 - \frac{1}{30}) = 1.933$ (1160 generations) and the mean time to a common ancestor of the population is $2(1 - \frac{1}{600}) = 1.997$ (1198 generations). The mean difference

between the time for a sample of size 30 to reach a MRCA, and the time for the whole population to reach its MRCA is 0.063, which is about 38 generations.

Warning The above calculations are not based on any of the basepair sequence information in the Nuu-Chah data set. They can only be viewed as crude guesses as to what one might expect from an unstructured randomly mating population. We will see later that our predictions can be refined once we fit the data to the model.

Note that T_2 makes a substantial contribution to the sum in (12) for T_{MRCAS}. For example, on average for over half the time since its MRCA, the sample will have exactly two ancestors.

Further, using independence of the T_k,

$$\begin{aligned}\mathrm{Var}(T_{\mathrm{MRCAS}}) &= \sum_{k=2}^{n} \mathrm{Var}(T_k) \\ &= \sum_{k=2}^{n} \left(\frac{2}{k(k-1)}\right)^2 \\ &= 8\sum_{k=1}^{n-1} \frac{1}{k^2} - 4\left(1-\frac{1}{n}\right)\left(3+\frac{1}{n}\right).\end{aligned}$$

It follows that

$$1 = \mathrm{Var}(T_2) \leq \mathrm{Var}(T_{\mathrm{MRCAS}}) \leq \lim_{n\to\infty} \mathrm{Var}(T_{\mathrm{MRCAS}}) = 8\frac{\pi^2}{6} - 12 \approx 1.16.$$

Exercise 5 Calculate the mean and standard deviation of the time to the MRCA of a population of $N = 600$. Express your answer in units of generations.

Lineage Sorting–an application in phylogenetics

Now focus on two particular individuals in the sample and observe that if these two individuals do not have a common ancestor at t, the whole sample cannot have a common ancestor. Since the two individuals are themselves a random sample of size two from the population, we see that

$$P(T_{\mathrm{MRCAS}} > t) \geq P(T_2 > t) = e^{-t},$$

it can be shown that

$$P(T_{\mathrm{MRCAS}} > t) \leq \frac{3(n-1)}{n+1} e^{-t} \tag{13}$$

and so

$$e^{-t} \leq P(T_{\text{MRCAS}} > t) \leq 3e^{-t} \tag{14}$$

The coalescent provides information on the history of genes within a population or species; by contrast, phylogenetic analysis studies the relationship between species. Central to a phylogenetic analysis of molecular data is the assumption that all individuals within a species have coalesced to a common ancestor at a more recent time point than the time of speciation, see Figure 6 for an illustration. If this assumption is met then it does not matter which homologous DNA sequence region is analyzed to infer the ancestral relationship between species. The true phylogeny should be consistently preserved regardless of the genetic locus used to infer the ancestry. If there is a discrepancy between the inferred phylogeny at one locus versus another then that discrepancy can be explained by the stochastic nature of statistical inference. However, the within species ancestry and the between species ancestry are not always on different time scales and completely separable. It is possible that a particular homologous region of DNA used to produce a phylogeny between species could produce a different phylogeny than a different homologous region and the difference is real (see Figure 7). One explanation of this phenomena is called *lineage sorting* and it occurs when the time to speciation is more recent than the time to the most recent common ancestry of the gene. This makes it appear like two subpopulations from the same species are more distantly related than two distinct species.

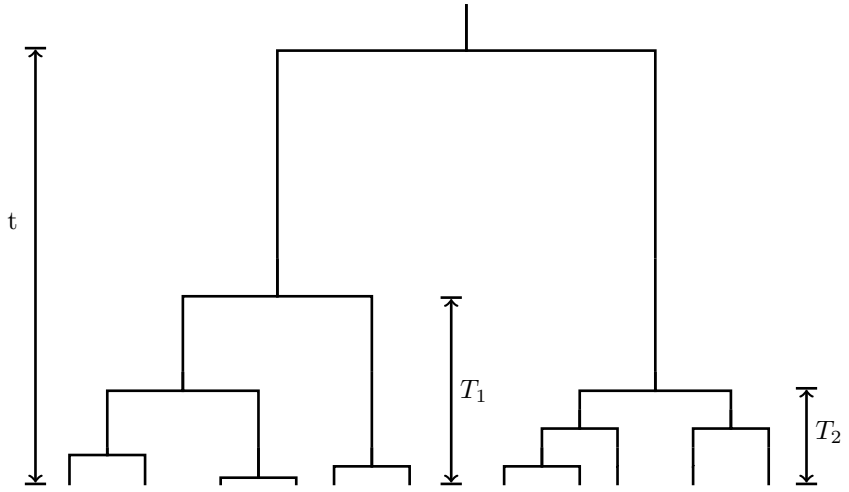

Figure 6: t: time to speciation, T_1: Time to common ancestor for population 1, T_2: Time to common ancestor for population 2. Population coalescence does not predate speciation.

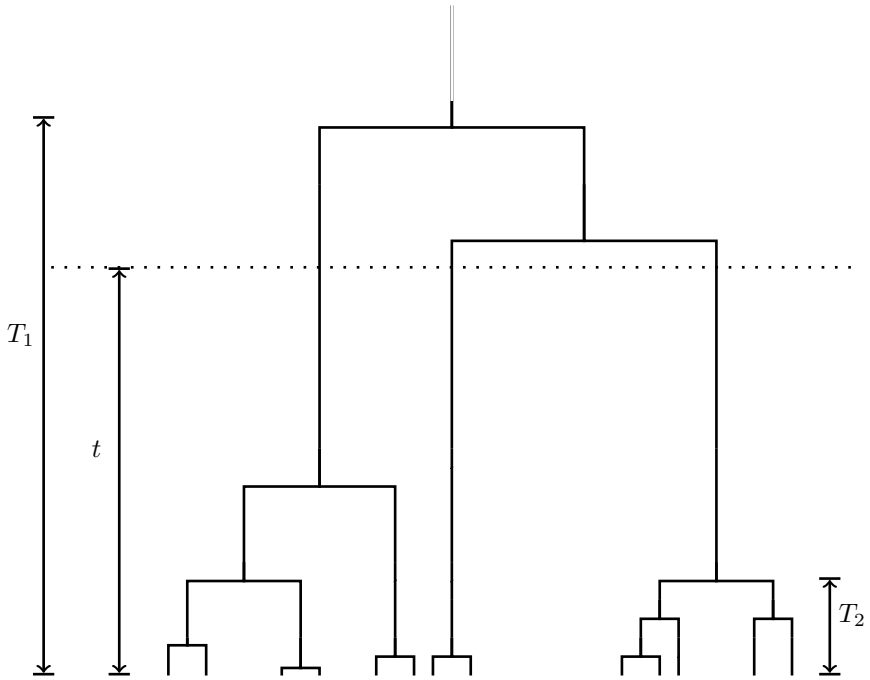

Figure 7: Population coalescence predates speciation.

However, the coalescent model can actually help determine if lineage sorting is plausible. For example, if based on external evidence, (possibly fossil evidence) the time to speciation is at least u generations into the past, then it is reasonable to ask, how likely is it that a population has not reached a common ancestor by time u. Converting from generations to coalescent the time scale, define $t = u/2N_e$. If T is the time it takes a population to reach a common ancestor, then we can use equation (14) to determine if lineage sorting is a reasonable explanation. If $3e^{-t}$ is small, then coalescent time scale and the phylogenetic time scales are likely to be different and lineage sorting is likely not to be the appropriate explanation. Thus another implication of coalescent theory is that the it provides appropriate insight as to how distantly related genes are within a species, which can help resolve issues in phylogenetic analysis.

The Effects of Variable Population Size

We study the evolution of a population whose total size varies over time. For convenience, suppose the model evolves in discrete generations and label the current generation as 0. Denote by $M(t)$ the number of haploid individuals in the population t generations before the present. To avoid pathologies, we

assume that $M(t) > 1$ for all t.

We assume that the variation in population size is due to either external constraints *e.g.* changes in the environment, or random variation which depends only on the total population size. This excludes so-called density dependent cases in which the variation depends on the genetic composition of the population, but covers many other settings. We continue to assume neutrality and random mating.

The usual approach to variable population size models is to approximate the evolution of the population with variable size by one with an appropriated constant size. There are various definitions of the appropriate constant or *effective population size* N_e; for a fixed time period $[0, T]$ is defined so that both populations will have the same reduction in heterozygosity:

$$\left(1 - \frac{1}{N_e}\right)^R = \prod_{r=1}^{R} \left(1 - \frac{1}{M(r)}\right)$$

and for large $M(r)$ via the same approximation used in **Exercise 3.2**

$$\frac{1}{N_e} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{M(r)}.$$

This last choice of N_e is the harmonic mean over $[1, T]$ of the population size. Among the disadvantages of this approach is the fact that the definition of N_e depends on a fixed time T, whereas many quantities depend on the evolution up to some random time.

We shall develop the theory of coalescent for a particular class of population growth models in which, roughly speaking, all the population sizes are large. Time will be scaled in units of $N = M(0)$ generations. To this end, define the relative size function $f_N(x)$ by

$$\begin{aligned} f_N(x) &= \frac{M(\lfloor Nx \rfloor)}{N} \\ &= \frac{M(r)}{N}, \qquad \tfrac{r}{N} \le x < \tfrac{r+1}{N}, \quad r = 0, 1, \ldots . \end{aligned} \tag{15}$$

We are interested in the behavior of the process when the size of each generation is large, so we suppose that

$$\lim_{N \to \infty} f_N(x) = f(x)$$

exists and is strictly positive for all $x \ge 0$.

Example 1 Many demographic scenarios can be modelled in this way. For an example of geometric population growth, suppose that for some constant $\rho > 0$ then

$$M(r) = \lfloor N(1 - \rho/N)^r \rfloor.$$

Then

$$\lim_{N\to\infty} f_N(x) = e^{-\rho x} \equiv f(x), x > 0.$$

Example 2 A commonly used model is one in which the population has constant size prior to generation V, and geometric growth from then to the present time. Thus for some $\alpha \in (0,1)$ then

$$M(r) = \begin{cases} \lfloor N\alpha \rfloor, & r \geq V \\ \lfloor N\alpha^{r/V} \rfloor, & r = 0, \ldots, V \end{cases}$$

If we suppose that $V = \lfloor Nv \rfloor$ for some $v > 0$, so that the expansion started v coalescent time ago, then

$$f_N(x) \to f(x) = \alpha^{\min(x.v,1)}.$$

The Coalescent in a Varying Environment

Consider first the Wright-Fisher model of reproduction, and note that the probability that two individuals chosen at time 0 have distinct ancestors r generations ago is

$$P(T_{N2} > r) = \prod_{j=1}^{r} \left(1 - \frac{1}{M(j)}\right),$$

where T_{N2} denotes the time to the common ancestor of the two individuals and $M(0) = N$. Recalling the inequality

$$x \leq -\log(1-x) \leq \frac{x}{1-x} < 1,$$

we see that

$$\sum_{j=1}^{r} \frac{1}{M(j)} \leq \sum_{j=1}^{r} \log\left(1 - \frac{1}{M(j)}\right) \leq \sum_{j=1}^{r} \frac{1}{M(j)-1}.$$

Rescaling time so that one unit of time corresponds to $N = M(0)$ generations, it follows that

$$\lim_{N\to\infty} -\sum_{j=1}^{\lfloor Nt \rfloor} \log\left(1 - \frac{1}{M(j)}\right) = \lim_{N\to\infty} \sum_{j=1}^{\lfloor Nt \rfloor} \frac{1}{M(j)}.$$

Since

$$\sum_{j=1}^{\lfloor Nt \rfloor} \frac{1}{M(j)} = \int_0^{(\lfloor Nt \rfloor - 1)/N} \frac{dx}{f_N(x)},$$

then

$$\lim_{N\to\infty} P(T_{N2} > \lfloor Nt \rfloor) = \exp\left(-\int_0^t \lambda(u)du\right),$$

where $\lambda(\cdot)$ is the coalescent intensity function defined by

$$\lambda(u) = \frac{1}{f(u)}, u \geq 0.$$

It is convenient to define

$$\Lambda(t) = \int_0^t \lambda(u)du,$$

the integrated intensity function. In the coalescent time scale with $T_{N2}/N \to T_2$ as $N\infty$ we have

$$P(T_2 > t) = \exp(-\Lambda(t)), t \geq 0.$$

Example 3 Return to the geometric growth model in **Example 1**. Note that

$$f(u) = e^{-\rho u}$$

implying $\lambda(u) = e^{\rho u}$ and

$$\Lambda(t) = \int_0^t e^{\rho u}du = .$$

Therefore,

$$P(T_2 > t) = \exp(-\frac{1}{\rho}(e^{\rho t} - 1).$$

For instance, under a fixed population size, the chance that it takes more than 1 unit of time (N generations) for two individuals to coalesce is $e^{-1} \approx .37$. Under a model with geometric growth

$$P(T_2 > 1) = \exp(-\frac{1}{\rho}(e^{\rho} - 1).$$

Now if $\rho = 2$ then $P(T_2 > 1) = .041$. Therefore, under population growth concrescences are occuring at a much faster rate, thus it is 9 times more rare for two individuals to take more than one unit of time to coalesce under a geometrically increasing population ($\rho = 2$) compared to the fixed population size model ($\rho = 0$).

Algorithm for generating coalescent times for varying population size model

In the last section we discussed the distribution of T_2 the time until two individuals coalesce under a varying population size model. For a sample of n individuals it is possible to derive the distribution of T_j the time the sample spends with j distinct ancestors. Unlike the fixed population case, the distribution of T_j will now depend on the time it takes to get to j ancestors. Define this time to be

$$S_j = T_n + T_{n-1} + \ldots, T_j$$

It can be shown that

$$P(T_j > t | S_{j+1} = s) = \exp\left(-\binom{j}{2}(\Lambda(s+t) - \Lambda(s))\right)$$

The above distribution can be turned into an algorithm for generating coalescent times under varying population size model. The algorithm goes as follows.

Algorithm for generating coalescent times $T_n, \ldots, T_2$

1. Generate $U_1, U_2, \ldots, U_n$ uniformly distributed random numbers.
2. Set $t = 0, \quad j = n$
3. Let $t_j^* = -\frac{2\log(U_j)}{j(j-1)}$
4. Solve for s in the equation
$$\Lambda(t+s) - \Lambda(t) = t_j^*$$
5. Set
$$t_j = s$$
$$t = t + s$$
$$j = j - 1$$
If $j \geq 2$ go to 3. Else return $T_n = t_n, \ldots, T_2 = t_2$

Note that t_j^* generated in step 2 above has an exponential distribution with parameter $j(j-1)/2$. If the population size is constant then $\Lambda(t) = t$, and so $t_j = t_j^*$

The Coalescent with mutation

We now introduce mutation, and suppose that the probability that any gene mutates in the time interval $(\tau + \delta\tau, \tau)$ is $(\theta/2)\delta\tau$. All mutants are assumed to be of new allelic types. Following the coalescent paradigm, we trace back the ancestry of a sample of n genes to the mutation forming the oldest allele in the sample. As we go backward in time along the coalescent, we shall encounter from time to time a "defining event", taken either as a coalescence of two lines of ascent into a common ancestor or a mutation in one or other of the lines of ascent. Figure 8 describes such an ancestry, identical to that of Figure 5 but with crosses to indicate mutations.

We exclude from further tracing back any line in which a mutation occurs, since any mutation occurring further back in any such line does not appear in the sample. Thus any such line may be thought of as stopping at the mutation, as shown in Figure 9 (describing the same ancestry as that in Figure 8).

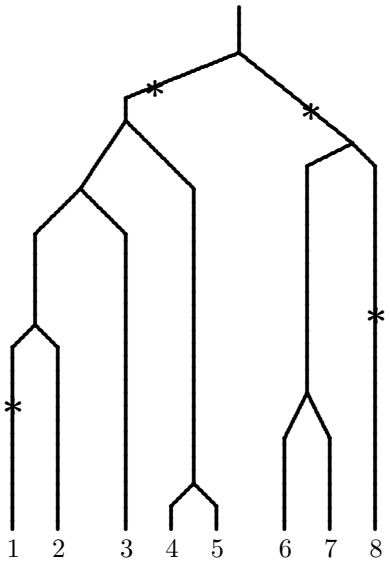

Figure 8: The coalescent with mutations

If at time τ there are j ancestors of the n genes in the sample, the probability that a defining event occurs in $(\tau, \tau + \delta\tau)$ is

$$\frac{1}{2}j(j-1)\delta\tau + \frac{1}{2}j\theta\delta\tau \;=\; \frac{1}{2}j(j+\theta-1)\delta\tau, \tag{16}$$

the first term on the left-hand side arising from the possibility of a coalescence of two lines of ascent, and the second from the possibility of a mutation.

If a defining event is a coalescence of two lines of ascent, the number of lines of ascent clearly decreases by 1. The fact that if a defining event arises from a mutation we exclude any further tracing back of the line of ascent in which the mutation arose implies that the number of lines of ascent also decreases by 1. Thus at any defining event the number of lines of ascent considered in the tracing back process decreases by 1. Given a defining event leading to j genes in the ancestry, the Poisson process theory described above shows that, going backward in time, the mean time until the next defining event occurs is $2/\{j(j+\theta-1)\}$, and that the same mean time applies when we restrict attention to those defining events determined by a mutation.

Thus starting with the original sample and continuing up the ancestry until the mutation forming the oldest allele in the sample is reached, we find that the mean age of the oldest allele in the sample is

$$2\sum_{j=1}^{n}\frac{1}{j(j+\theta-1)} \tag{17}$$

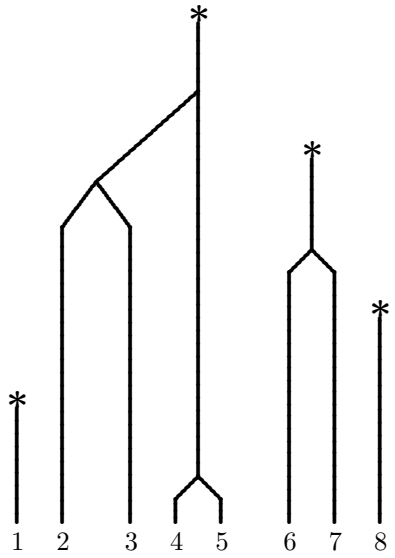

Figure 9: Tracing back to, and stopping at, mutational events

coalescent time units. The value in (17) must be multiplied by $2N$ to give this mean age in terms of generations.

The time backward until the mutation forming the oldest allele in the sample, whose mean is given in (17), does not necessarily trace back to, and past, the most recent common ancestor of the genes in the sample (MRCAS), and will do so only if the allelic type of the MRCAS is represented in the sample. This observation can be put in quantitative terms by comparing the MRCAS given in (12) to the expression in (17). For small θ, the age of the oldest allele will tend to exceed the time back to the MRCAS, while for large θ, the converse will tend to be the case. The case $\theta = 2$ appears to be a borderline one: For this value, the expressions in (12) and (17) differ only by a term of order n^{-2}. Thus for this value of θ, we expect the oldest allele in the sample to have arisen at about the same time as the MRCAS.

The competing Poisson process theory outlined above shows that, given that a defining event occurs with j genes present in the ancestry, the probability that this is a mutation is $\theta/(j-1+\theta)$. Thus the mean number of different allelic types found in the sample is

$$\sum_{j=1}^{n} \frac{\theta}{j-1+\theta}.$$

The number of "mutation-caused" defining events with j genes present in the ancestry is, of course, either 0 or 1, and thus the variance of the number of

different allelic types found in the sample is

$$\sum_{j=1}^{n} \left(\frac{\theta}{j-1+\theta} - \frac{\theta^2}{(j-1+\theta)^2} \right).$$

Even more than this can be said. The probability that exactly k of the defining events are "mutation-caused" is clearly proportional to $\theta^k/\{\theta(\theta+1)\cdots(\theta+n-1)\}$, the proportionality factor not depending on θ.

The sample contains only one allele if no mutants occurred in the coalescent after the original mutation for the oldest allele. Moving up the coalescent, this is the probability that all defining events before this original mutation is reached are amalgamations of lines of ascent rather than mutations. The probability of this is

$$\prod_{j=1}^{n-1} \frac{j}{(j+\theta)} \;=\; \frac{(n-1)!}{(1+\theta)(2+\theta)\cdots(n-1+\theta)}. \tag{18}$$

Allele frequencies and site frequencies

Introduction

The current direction of interest in population genetics is a retrospective one, looking backwards to the past rather than looking forward into the future. This change of direction is largely spurred by the large volume of genetic data now available at the molecular level and a wish to infer the forces that led to the data observed. This data driven modern perspective will be the focus of the notes that follow.

The material in this section covers both sample and population formulas relating to the infinitely many alleles model.

Allele frequencies

We first discuss allelic frequencies, for which finding "age" properties amounts to finding size-biased properties. Kingman's (1975) Poisson–Dirichlet distribution, which arises in various allelic frequency calculations, is not user-friendly. This makes it all the more interesting that a *size-biased* distribution closely related to it, namely the GEM distribution, named for Griffiths, (1980), Engen (1975) and McCloskey (1965), who established its salient properties, is both simple and elegant. More important, it has a central interpretation with respect to the ages of the alleles in a population. We now describe this distribution.

Suppose that a gene is taken at random from the population. The probability that this gene will be of an allelic type whose frequency in the population is x is just x. In other words, alleles are sampled by this choice in a size-biased way. The probability that there exists an allele in the population with frequency between x and $x + \delta x$. It follows that the probability the gene chosen is of

this allelic type is $\theta x^{-1}(1-x)^{\theta-1}x\delta x = \theta(1-x)^{\theta-1}\delta x$. From this, the density function $f(x)$ of the frequency of this allele is given by

$$f(x) = \theta(1-x)^{\theta-1}. \tag{19}$$

Suppose now that all genes of the allelic type just chosen are removed from the population. A second gene is now drawn at random from the population and its allelic type observed. The frequency of the allelic type of this gene among the genes remaining at this stage can be shown to also be given by (19). All genes of this second allelic type are now also removed from the population. A third gene then drawn at random from the genes remaining, its allelic type observed, and all genes of this (third) allelic type removed from the population. This process is continued indefinitely. At any stage, the distribution of the frequency of the allelic type of any gene just drawn among the genes left when the draw takes place is given by (19). This leads to the following representation. Denote by w_j the original population frequency of the jth allelic type drawn. Then we can write $w_1 = x_1$, and for $j = 2, 3, \ldots$,

$$w_j = (1-x_1)(1-x_2)\cdots(1-x_{j-1})x_j, \tag{20}$$

where the x_j are independent random variables, each having the distribution (19). The random vector $(w_1, w_2, \ldots)$ then has the GEM distribution.

All the alleles in the population at any time eventually leave the population, through the joints processes of mutation and random drift, and any allele with current population frequency x survives the longest with probability x. That is, since the GEM distribution was found according to a size-biased process, it also arises when alleles are labeled according to the length of their future persistence in the population. Reversibility arguments then show that the GEM distribution also applies when the alleles in the population are labeled by their age. In other words, the vector $(w_1, w_2, \ldots)$ can be thought of as the vector of allelic frequencies when alleles are ordered with respect to their ages in the population (with allele 1 being the oldest).

The elegance of many age-ordered formulae derives directly from the simplicity and tractability of the GEM distribution. We now give two examples. First, the GEM distribution shows immediately that the mean population frequency of the oldest allele in the population is

$$\theta \int_0^1 x(1-x)^{\theta-1} = 1/(1+\theta), \tag{21}$$

and more generally that the mean population frequency of the jth oldest allele in the population is

$$\frac{1}{1+\theta}\Big(\frac{\theta}{1+\theta}\Big)^{j-1}.$$

Second, the probability that a gene drawn at random from the population is of the type of the oldest allele is the mean frequency of the oldest allele, namely

$1/(1+\theta)$, as just shown. More generally the probability that n genes drawn at random from the population are all of the type of the oldest allele is

$$\theta \int_0^1 x^n (1-x)^{\theta-1}\, dx = \frac{n!}{(1+\theta)(2+\theta)\cdots(n+\theta)}.$$

The probability that n genes drawn at random from the population are all of the same unspecified allelic type is

$$\theta \int_0^1 x^{n-1} (1-x)^{\theta-1}\, dx = \frac{(n-1)!}{(1+\theta)(2+\theta)\cdots(n+\theta-1)}.$$

From this, given that n genes drawn at random are all of the same allelic type, the probability that they are all of the allelic type of the oldest allele is $n/(n+\theta)$.

Ages

We turn now to sample properties, which are in practice more important than population properties. The most important sample distribution concerns the frequencies of the alleles in the sample when ordered by age. This distribution was found by Donnelly and Tavaré (1986), who showed that the probability that the number K_n of alleles in the sample takes the value k, and that the age-ordered numbers of these alleles in the sample are, in age order, $n_{(1)}, n_{(2)}, \ldots, n_{(k)}$, is

$$\frac{\theta^k (n-1)!}{S_n(\theta) n_{(k)} (n_{(k)} + n_{(k-1)}) \cdots (n_{(k)} + n_{(k-1)} + \cdots n_{(2)})}, \tag{22}$$

where $S_n(\theta)$ is defined

$$S_n(\theta) = \theta(\theta+1)(\theta+2)\cdots(\theta+n-1). \tag{23}$$

.

Several results concerning the oldest allele in the sample can be found from this formula, or in some cases more directly by other methods. For example, the probability that the oldest allele in the sample is represented by j genes in the sample is (Kelly, (1976))

$$\frac{\theta}{n} \binom{n}{j} \binom{n+\theta-1}{j}^{-1}. \tag{24}$$

Further results provide connections between the oldest allele in the sample to the oldest allele in the population. Some of these results are exact for a Moran model and others are the corresponding diffusion approximations. For example, Kelly (1976) showed that in the Moran model, the probability that the oldest allele in the population is observed at all in the sample is $n(2N+\theta)/[2N(n+\theta)]$. This is equal to 1, as it must be, when $n = 2N$, and for the case $n = 1$ reduces to a result found above that a randomly selected gene is of the oldest allelic

type in the population. The diffusion approximation to this probability, found by letting $N \to \infty$, is $n/(n+\theta)$.

A further result is that in the Moran model, the probability that a gene seen j times in the sample is of the oldest allelic type in the population is $j(2N+\theta)/[2N(n+\theta)]$. Letting $N \to \infty$, the diffusion approximation for this probability is $j/(n+\theta)$. When $n = j$ this is $j/(j+\theta)$, a result found above found by other methods.

Donnelly (1986)) provides further formulae extending these. He showed, for example, that the probability that the oldest allele in the population is observed j times in the sample is

$$\frac{\theta}{n+\theta}\binom{n}{j}\binom{n+\theta-1}{j}^{-1}, \quad j = 0, 1, 2, \ldots, n. \tag{25}$$

This is of course closely connected to the Kelly result (24). For the case $j = 0$ this probability is $\theta/(n+\theta)$, confirming the complementary probability $n/(n+\theta)$ found above. Conditional on the event that the oldest allele in the population does appear in the sample, a straightforward calculation using (25) shows that this conditional probability and that in (24) are identical.

Griffiths and Tavaré (1998) give the Laplace transform of the distribution of that age of an allele observed b times in a sample of n genes, together with a limiting Laplace transform for the case when θ approaches 0. These results show, for the Wright–Fisher model, that the diffusion approximation for the mean age of such an allele is

$$\sum_{j=1}^{\infty} \frac{4N}{j(j-1+\theta)}\Big(1 - \frac{(n-1-b+\theta-1)_{(j)}}{(n-1+\theta-1)_{(j)}}\Big) \tag{26}$$

generations, where $a_{(j)}$ is defined as $a_{(j)} = a(a+1)\cdots(a+j-1)$.

Site (SNP) Frequencies

The length of a coalescent tree is defined to be the sum of all of its branch lengths and is denoted by L_n which can be determined from the coalescent times as follows

$$L_n = \sum_{j=2}^{n} jT_j,$$

where the random variable T_j are independent and have exponential distribution with rate parameter $j(j-1)/2$. If S_n denotes the total number of mutations on the genealogical tree back to the MRCA of a sample of size n, then conditional

on L_n, S_n has a Poisson distribution with mean $\theta L_n/2$. It follows that

$$\begin{aligned} E(S_n) &= E(E(S_n|L_n)) \\ &= E(\theta L_n/2) \\ &= \frac{\theta}{2} E(\sum_{i=2}^{n} iT_i) \\ &= \frac{\theta}{2} \sum_{i=2}^{n} iE(T_i) \\ &= \frac{\theta}{2} \sum_{i=2}^{n} i \frac{2}{i(i-1)} \\ &= \theta \sum_{j=1}^{n-1} \frac{1}{j} \end{aligned} \tag{27}$$

Notice that for large n then $E(S_n) \sim \theta \log n$.

Example 3 We calculate the mean number of mutations for various sample sizes, when $\theta = 4$,

Sample size n	θ	$E(S_n)$
10	4	9.21
20	4	11.98
40	4	14.76
45	4	15.23
50	4	15.65
60	4	16.38
100	4	18.42

Recall that $E(S_n)$ is the average number of mutations accumulated by a sample of size n under the neutral coalescent model. Any given realization of evolution will produce an S_n that varies around the expected value. The standard deviation of S_n tells you how much variation to expect. The formula for standard deviation is $\text{STDEV}(S_n) = \sqrt{\text{Var}(S_n)}$. We will now calculate $\text{Var}(S_n)$. Because S_n arises as a result of two random processes, we need to account for both processes in our calculation of the variance of S_n. Below is the formula that is needed to calculate $\text{Var}(S_n)$.

$$\text{Var}(S_n) = E(\text{Var}(S_n|L_n)) + \text{Var}(E(S_n|L_n)). \tag{28}$$

One way to interpret Equation (28) is as follows. There are two sources of variation. One is due to fluctuations inherent in the coalescent process and the

other is due to the fluctuations inherent in the Poisson mutation process. The term $E(\text{Var}(S_n|L_n))$ can be thought as the contribution of the variance of S_n attributed to the Poisson mutation process. $\text{Var}(E(S_n|L_n))$ may be though of as the amount variation due to the coalescent process. Therefore,

$$\begin{aligned}
\text{Var}(S_n) &= E(\text{Var}(S_n|L_n)) + \text{Var}(E(S_n|L_n)) \\
&= E\left(\frac{\theta}{2}L_n\right) + \text{Var}\left(\frac{\theta}{2}L_n\right) \\
&= \frac{\theta}{2}E(L_n) + \frac{\theta^2}{4}\text{Var}(L_n) \\
&= \theta\sum_{i=1}^{n-1}\frac{1}{i} + \frac{\theta^2}{4}\text{Var}\left(\sum_{i=2}^{n} iT_i\right) \\
&= \theta\sum_{i=1}^{n-1}\frac{1}{i} + \frac{\theta^2}{4}\sum_{i=2}^{n} i^2\text{Var}\left(T_i\right) \\
&= \theta\sum_{i=1}^{n-1}\frac{1}{i} + \frac{\theta^2}{4}\sum_{i=2}^{n} i^2\frac{4}{i^2(i-1)^2} \\
&= \theta\sum_{i=1}^{n-1}\frac{1}{i} + \theta^2\sum_{i=1}^{n-1}\frac{1}{i^2}
\end{aligned}$$

For large n then

$$\text{Var}(S_n) = \theta\log n + 2\theta^2$$

Below is a table means and standard deviations

n	θ	$E(S_n)$	$\text{Stdev}(S_n)$
10	4	9.21	6.00
20	4	11.98	6.10
40	4	14.76	6.20
45	4	15.23	6.21
50	4	15.65	6.23
60	4	16.38	6.25
100	4	18.42	6.32

Estimating the parameter θ

We have been investigating the properties of the neutral coalescent. We have been focusing our efforts on answering the following question: if the neutral model for evolution with constant mutation rate is a reasonable model, what can we expect the ancestry of a sample to look like? We found that under

neutrality coalescence occur at the rate of $n(n-1)/2$ where n is the sample size. This means that on average, coalescence occur quickly in the recent past and then very slowly in the more distant past, as the number of ancestors becomes small. In fact, on average, half the time back to MRCA is T_2 the time for last two ancestors to coalesce. We found that the average number of mutations back to the MRCA is proportional to the mutation parameter θ and inversely proportional to $\log n$.

Of course, averages tell only part of the story. There is a fair amount of variation about the average. To get a handle on the variation, we calculated the variance and standard deviation for T_n, the time back to the MRCA, and the variance and standard deviation of S_n, the number of mutations back to the MRCA of a sample of size n.

We now want to shift the focus from mathematical modelling to statistical inference. Rather than ask, 'for a given mutation parameter, θ, what can we say about the ancestry of the sample, we now ask the more relevant question, for a given sample, what can we say about the population. In particular, what is our best estimate for θ based on information in a sample.

Watterson's estimator

Under the assumptions of the infinite sites model, the number of segregating sites is exactly the total number of mutations S_n since the MRCA of the sample. Recall that

$$E(S_n) = a_n\theta \tag{29}$$

where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ and

$$\mathrm{Var}(S_n) = a_n\theta + b_n\theta^2 \tag{30}$$

where $b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}$.

Define

$$\hat{\theta}_S = \frac{S_n}{a_n}.$$

This is called the segregating sites estimator for θ and goes back to a paper by Watterson (1975). Note that it follows from (29) that $E(\hat{\theta}_S) = \theta$. Estimators of this type are called unbiased. It follows from (30) that

$$\mathrm{Var}(\hat{\theta}_S) = \frac{1}{a_n}\theta + \frac{b_n}{a_n^2}\theta^2 \tag{31}$$

It is easy to see that $\mathrm{Var}(\hat{\theta}_S) \to 0$ as $n \to \infty$. Estimators with this property are said to be consistent. This means that one can attain any level of precision desired by choosing the sample size sufficiently large. However, don't expect

the precision to be much better than half the size of the estimate, unless you require ridiculously large sample sizes.

Example 4 If it is known that θ is no bigger than 6, using the segregating sites estimator for θ, how large a sample is required to insure that the error of the estimate is less than or equal to 1?
Soln. Lets assume that the error of an estimate is 2 standard deviations. If 2 standard deviations is 1 unit, then we want to choose a sample size so that the standard deviation of the estimate for θ is less than or equal to .5. Using a conservative initial guess for θ to be 6 we have

$$.5 = \sqrt{\frac{6}{a_n} + \frac{36b_n}{a_n^2}}$$

We wish to solve for n in the above equation. To simplify matters lets replace b_n with its upper bound of 2. Therefore

$$\begin{aligned} .25 &= \frac{6}{a_n} + \frac{72}{a_n^2} \\ .25a_n^2 &= 6a_n + 72 \end{aligned}$$

Solving the above quadratic equation gives $a_n \approx 32$, implying $n \approx 1.73 \times 10^{14}$. However, if you require a standard deviation for the estimate to be 2, then the sample size required for this level of precision is just $n = 158$.

Pairwise differences

Recall that θ is the expected number of mutations separating two individuals. So a natural way to estimate θ is to calculate number of mutations separating individuals two at a time and average over all pairs. This may be thought of as a sample average used to estimate a population average. To calculate this we take individuals two at a time. Denote by

$$S_{ij} = \text{Number of mutations separating individuals i and j.}$$

Under the infinite sites assumption, we can calculate S_{ij} from a sample by calculating the number of segregating sites between sequences i and j. If we average S_{ij} over all pairs (i, j), this is called the average number of pairwise differences. We denote the average number of pairwise differences by.

$$D_n = \frac{2}{n(n-1)} \sum_{i \leq j} S_{ij}.$$

Note that we can think of individuals (i, j) as sample of size 2, therefore

$$E(S_{ij}) = E(S_2) = \theta.$$

Therefore,

$$E(D_n) = \frac{2}{n(n-1)} \sum_{i \leq j} E\left(S_{ij}\right) = \theta$$

Thus, D_n is an unbiased estimator. Tajima (1981) was the first to investigate the properties of D_n. We will refer to $\hat{\theta}_T = D_n$. It is interesting to note that $\hat{\theta}_T$ has very poor statistical properties. In fact, $\hat{\theta}_T$ has higher variance than any of the other estimators we will consider. Why does an estimator that seems so natural have such poor properties? The answer lies in the fact that their is dependence in the data generated by the common ancestral history. This means that S_{ij} and S_{kl} are positively correlated random variables. As a result the precision of the estimator D_n will be low.

In fact,

$$\mathrm{Var}(D_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \tag{32}$$

The details for deriving $\mathrm{Var}(D_n)$ are left as an exercise (se below)

Note that

$$\lim_{n\to\infty} \mathrm{Var}(D_n) = \frac{\theta}{3} + \frac{2}{9}\theta^2.$$

The pairwise difference estimate is not consistent. The square root of the above limit represents the optimal precision one can obtain, regardless of sample size, using the pairwise difference estimator.

Exercise 6

1. Show that

$$\begin{aligned} E(D_n^2) &= \frac{1}{n^2(n-1)^2}\left[2n(n-1)E(S_{12}^2) + 4n(n-1)(n-2)E(S_{12}S_{13})\right. \\ &\left. +n(n-1)(n-2)(n-3)E(S_{12}S_{34})\right]. \end{aligned}$$

2. Show that $E(S_{12}^2) = 2\theta^2 + \theta$.

3. It can be shown that

$$E(S_{12}S_{13}) = \frac{4\theta^2}{3} + \frac{\theta}{2},$$

and

$$E(S_{12}S_{34}) = \frac{11\theta^2}{9} + \frac{\theta}{3}.$$

Use these results to calculate $E(D_n^2)$

4. Show that

$$\mathrm{Var}(D_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$$

Estimating θ in the infinitely many alleles model

For the Wright–Fisher infinitely many alleles model, the conditional distribution of the vector $\mathbf{A} = (A_1, A_2, \ldots, A_n)$, given the value of K_n, is

$$\text{Prob}\{\mathbf{A} = \mathbf{a} | K_n = k\} = \frac{n!}{|S_n^k| \; 1^{a_1} 2^{a_2} \; \ldots \; n^{a_n} \, a_1! a_2! \ldots a_n!}\,, \tag{33}$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_n)$. This conditional distribution is exact for the Moran model, and we use it as the basis of the theory for estimating θ in infinitely many alleles models.

Equation (33) implies that K_n is a *sufficient statistic* for θ. Standard statistical theory then shows that, once the observed value k_n of K_n is given, no further information about θ is provided by the various a_j values, so that all inferences about θ should be carried out using the observed value k_n of K_n only. This includes estimation of θ or of any function of θ.

Since K_n is a sufficient statistic for θ we can find the maximum likelihood estimator $\hat{\theta}_K$ of θ. It is found that this estimator is the implicit solution of the equation

$$K_n = \frac{\hat{\theta}_K}{\hat{\theta}_K} + \frac{\hat{\theta}_K}{\hat{\theta}_K + 1} + \frac{\hat{\theta}_K}{\hat{\theta}_K + 2} + \cdots + \frac{\hat{\theta}_K}{\hat{\theta}_K + n - 1}\,. \tag{34}$$

Numerical calculation of the estimate $\hat{\theta}_k$ using (34) is usually necessary.

Likelihood and Efficiency

In the last section we considered two estimators for θ that were based on summary statistics. Noticed that the segregating sites method performed better than the pairwise difference method. However, both estimators tend to have fairly high variance. The theory of mathematical statistics provides us with a lower bound on the variance of all unbiased estimators. This lower bound is called the Cramèr-Rao lower bound. Efficiency of an estimator is defined to be the variance of an estimator relative to the minimum variance possible. In this section we begin with some general results from mathematical statistics. In particular we establish the Cramèr-Rao lower bound. We then calculate this lower bound in the context of the neutral coalescent model.

General Set up Let $X_1, X_2, \cdots, X_n$ be a sample of size n with

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1, x_2, \ldots, x_n; \theta).$$

We wish to estimate the parameter θ. An estimate of θ is a function of the data. Let $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, \ldots, X_n)$ be an estimator of θ. An unbiased estimator has the property that

$$E(\hat{\theta}) = \theta.$$

Result 1

$$E\left(\frac{\partial}{\partial\theta}\log f(X_1, X_2, \ldots, X_n; \theta)\right) = 0$$

Proof.

Define

$$u(x_1, x_2, \ldots, x_n; \theta) = \frac{\partial}{\partial\theta}\log f(x_1, x_2, \ldots, x_n; \theta)$$

which can also be written as

$$u(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{f(x_1, x_2, \ldots, x_n; \theta)}\frac{\partial}{\partial\theta} f(x_1, x_2, \ldots, x_n; \theta).$$

If we define a random variable $U = u(X_1, X_2, \ldots, X_n; \theta)$ then

$$\begin{aligned} E(U) &= \sum u(x_1, x_2, \ldots, x_n; \theta) f(x_1, x_2, \ldots, x_n) \\ &= \sum \frac{\partial}{\partial\theta} f(x_1, x_2, \ldots, x_n; \theta) \\ &= \frac{\partial}{\partial\theta} \sum f(x_1, x_2, \ldots, x_n; \theta) \\ &= \frac{\partial}{\partial\theta} 1 \\ &= 0 \end{aligned}$$

Result 2

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f(X_1, X_2, \ldots, X_n; \theta)\right) = -E\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1, X_2, \ldots, X_n; \theta)\right)$$

The proof is left as an exercise

Cramèr-Rao Lower Bound. If $\hat{\theta}$ is an unbiased estimator of θ then

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{-E\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1, X_2, \ldots, X_n; \theta)\right)}$$

Proof. Note that

$$\theta = E(\hat{\theta}) = \sum \hat{\theta}(x_1, x_2, \ldots, x_n) f(x_1, \ldots, x_n; \theta)$$

Differentiating the above equation with respect to θ gives

$$\begin{aligned} 1 &= \sum \hat{\theta}(x_1, x_2, \ldots, x_n) \frac{\partial}{\partial \theta} f(x_1, \ldots, x_n; \theta) \\ &= \sum \hat{\theta}(x_1, x_2, \ldots, x_n) u(x_1, x_2, \ldots, x_n; \theta) f(x_1, \ldots, x_n; \theta) \\ &= E(U\hat{\theta}) \\ &= \mathrm{Cov}(U, \hat{\theta}) \end{aligned}$$

The last line follows from the fact that Cov $(U, \hat{\theta}) = E(U\hat{\theta}) - E(U)E(\hat{\theta})$. Recall from Result 1 that $E(U) = 0$.

Because the correlation coefficient is always between ± 1, it follows that

$$\mathrm{Var}(\hat{\theta})\mathrm{Var}(U) \geq [\mathrm{Cov}(U, \hat{\theta})]^2.$$

Therefore

$$\mathrm{Var}(\hat{\theta})\mathrm{Var}(U) \geq 1$$

implying

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathrm{Var}(U)}.$$

It follows from Result 2 that $\mathrm{Var}(U) = -E\left(\frac{\partial^2}{\partial \theta^2} \log f(X_1, X_2, \ldots, X_n; \theta)\right)$.

Lower bound for the Variance of the mutation parameter θ

Suppose that we assume that every mutation that separates all individuals at a particular locus in the population is revealed, and the full ancestry is resolved. Further assume that the number of mutations between each coalescent event is observable. Define Y_j to be the number of mutations that occur during the time the sample has j distinct ancestors. Therefore, $P(Y_j = y_j)$ is the probability that y_j mutations occur before a coalescence. This is analogous to flipping an (unfair) coin and asking what is the probability of getting y_j tails before a heads. This produces the well known geometric distribution given by

$$P(Y_j = y_j) = \left(\frac{\theta}{j - 1 + \theta}\right)^{y_j} \left(\frac{j-1}{j-1+\theta}\right).$$

Because of independence we can write,

$$\begin{aligned} f(y_2, y_3, \ldots, y_n; \theta) &= P(Y_2 = y_2, Y_2 = y_3, \cdots, Y_n = y_n; \theta) \\ &= \prod_{j=2}^{n} P(Y_j = y_j) \\ &= \prod_{j=2}^{n} \left(\frac{\theta}{j - 1 + \theta}\right)^{y_j} \left(\frac{j-1}{j-1+\theta}\right) \end{aligned} \tag{35}$$

For notational convenience we will denote the likelihood by

$$L_n(\theta) = f(Y_2, Y_3, \ldots, Y_n; \theta)$$

It is easy to check that

$$\frac{\partial^2}{\partial\theta^2} \log L_n = -\frac{S_n}{\theta^2} + \sum_{j=2}^{n} \frac{Y_j + 1}{(j-1+\theta)^2}$$

so that

$$\begin{aligned} -E\left(\frac{\partial^2}{\partial\theta^2} \log L_n\right) &= \frac{\sum_1^{n-1} \frac{1}{j}}{\theta} - \sum_{j=2}^{n} \left(\frac{\theta}{j-1} + 1\right) \frac{1}{(j-1+\theta)^2} \\ &= \frac{1}{\theta} \sum_{j=1}^{n-1} \frac{1}{j} - \sum_{j=1}^{n-1} \frac{1}{j(j+\theta)} \\ &= \frac{1}{\theta} \sum_{j=1}^{n-1} \frac{1}{j} - \frac{1}{\theta} \sum_{j=1}^{n-1} \left(\frac{1}{j} - \frac{1}{j+\theta}\right) \\ &= \frac{1}{\theta} \sum_{j=1}^{n-1} \frac{1}{j+\theta} \end{aligned}$$

Hence the variance of any unbiased estimators $\hat{\theta}$ of θ satisfies

$$\mathrm{Var}(\hat{\theta}) \geq \frac{\theta}{\sum_{j=1}^{n-1} \frac{1}{j+\theta}} \equiv \mathrm{Var}(\hat{\theta}_F)$$

Note that $\sum_{j=1}^{n-1} 1/(\theta+j) \approx \log(\theta+n)$. So as n-the number of individuals in the sample becomes large, the variance of the estimator will decrease at a very slow rate. The above Cramér-Rao lower bound on the variance shows that among unbiased estimators the best one can do is this lower bound.

This result is due to Fu and Li (1993). We will refer to the optimal estimator of Fu and Li as $\hat{\theta}_F$. The standard deviation efficiency for the Watterson's segregating sites estimator $\hat{\theta}_S$ and the Tajima's pairwise differences estimator $\hat{\theta}_T$ is given by

$$\sqrt{\frac{\mathrm{Var}(\hat{\theta}_F)}{\mathrm{Var}(\hat{\theta}_S)}}$$

and

$$\sqrt{\frac{\mathrm{Var}\hat{\theta_F}}{\mathrm{Var}\hat{\theta_T}}}$$

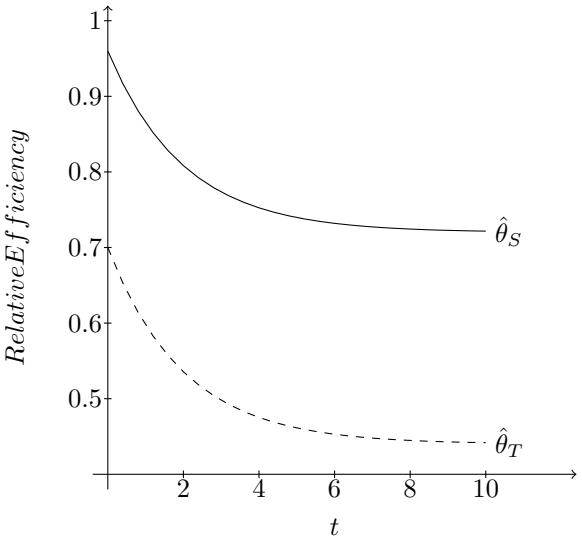

Figure 10: Relative efficiency of the pairwise differences estimator $\hat{\theta}_T$ (dashed) versus the relative efficiency of the segregating sites estimator $\hat{\theta}_S$ (solid). $0 < \theta < 10$ and $n = 50$.

respectively. What follows are some plots of the standard deviation relative efficiency for the pairwise difference and segregating sites estimators (see Figure 10).

While it is true that both the 'best' estimator and the segregating sites estimator have variance that converges to zero at rate $\log n$, the graphs in Figure 11 show that extremely large sample sizes are required before the segregating sites variance comes close to that of the optimal estimator. However, the Fu estimator is based on a likelihood (equation 35) that requires knowing the number of mutations between coalescent events. This is unobservable. To obtain a maximum likelihood estimate based on observed data, we need to consider a more computationally intensive approach. To gain some appreciation for the amount of computation required to implement a maximum likelihood approach, we begin by considering the full likelihood on a very small data set.

A numerical example using a small data set

Consider the following simple example. We have three sequences and four segregating sites and each sequence has multiplicity unity. Using the binary code discussed in the exercises we describe the data as follows.

$$\begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{matrix}$$

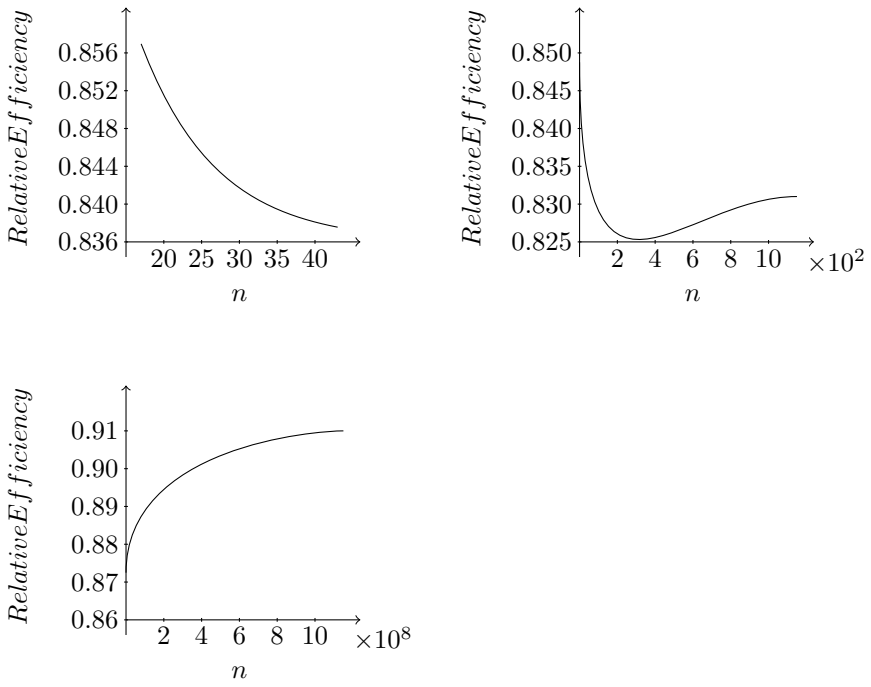

Figure 11: Relative efficiency of the segregating sites estimator $\hat{\theta}_S$ as a function of sample size. Small sample size (left top), moderate sample size (right top) and large sample size (left bottom).

For convenience, label the segregating sites 1,2,3 and 4 from left to right. There are five possible labeled rooted trees constructed from the unrooted genealogy. These five rooted gene trees for this data are shown in Figure 12. The possible coalescent trees producing Figure 12 are given in Figure 13.

Let T_3 be the time during which the sample has three ancestors, and T_2 the time during which it has two. By considering the Poisson nature of the mutations along the edges of the coalescent tree, the probability of each type of tree can be calculated.

For example, the probability $p_{(1a)}$ of the first labelled tree (a) is

$$\begin{aligned} p_{(a1)} &= E\left[\left(e^{-\theta T_3/2}\frac{\theta T_3}{2}\right)^2 e^{-\theta T_2/2} e^{-\theta(T_2+T_3)/2}\frac{1}{2!}(\theta(T_2+T_3)/2)^2\right] \\ &= \frac{\theta^4}{32}E\left[e^{-\theta(3T_3/2+T_2)}T_3^2(T_2+T_3)^2\right] \\ &= \frac{\theta^4(17\theta^2+46\theta+32)}{27(\theta+1)^3(\theta+2)^5} \end{aligned}$$

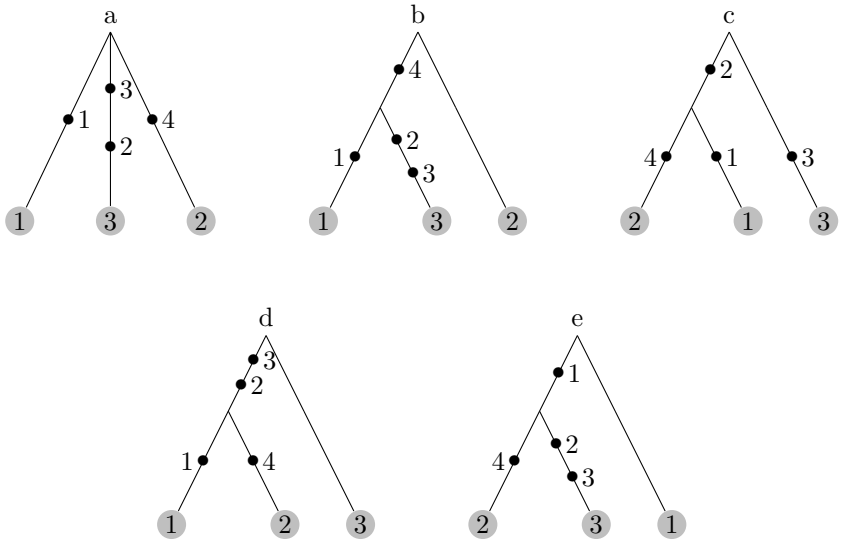

Figure 12: Gene trees consistent with the 4 segregating sites.

We must now do a similar calculation for each of the remaining five coalescent trees and sum the results. While it is indeed possible to calculate the likelihood explicitly for this extremely small data set, it is clear that a more feasible approach will be required for more realistic data sets. You can see that the number of coalescent trees consistent with the data will grow rapidly as we increase the number of sequences.

Computationally intensive methods

It is not an exaggeration to say that Markov Chain Monte Carlo (MCMC) methods have revolutionized statistics and are at the heart of many computationally intensive methods. So it may be surprising to note that the most commonly used MCMC method, called the Metropolis Hastings Algorithm, is only three lines of code and the mathematical argument that justifies its legitimacy is only four lines long. In fact, the ease at which one can produce an MCMC algorithm to address a particular statistical problem can be seen as a drawback. The simplicity of the algorithm often leads individuals to try MCMC as there first method toward a solution. However, MCMC should be the algorithm of last resort. If all else fails, use MCMC. The reason for this is that the MCMC algorithm is plagued with tricky convergence issues and requires extensive diagnostics before one can reliably trust the answer. However, even with all its potential drawbacks and pitfalls, it is still an incredibly useful tool in statistics.

Since a good deal of this course involves various types of Markov processes,

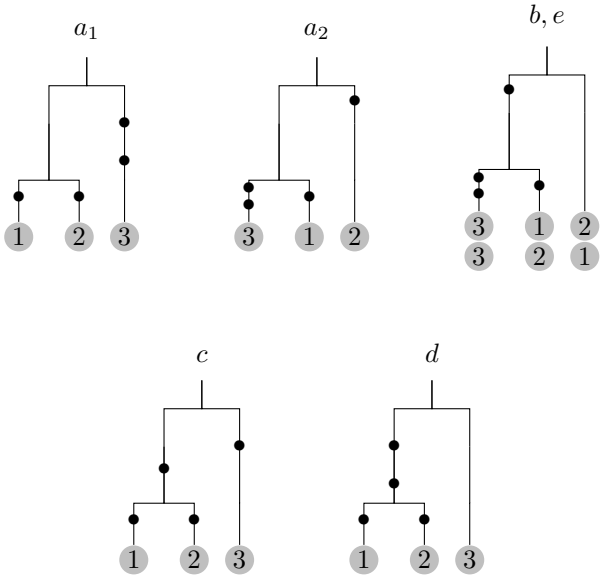

Figure 13: Coalescent trees consistent with the gene trees.

it is worth pointing out the distinction between the Markov processes discussed in detail by Dr. Ewens and Markov Chain Monte Carlo methods discussed here. The typical approach to stochastic mathematical modeling is to begin with a probabilistic description of the phenomena of interest. In much of this course we are concerned with how population factors effect genetic variation over evolutionary time. Examples of mathematical descriptions that address this problem include the Moran Model, the Wright-Fisher Model and the general Cannings Model. These are all Markov models. Within the context of these models we are interested in long term behavior which often leads to a stationary distribution of the process of interest. The natural progression of ideas starts with a Markov model and from this we derive the stationary distribution.

However, MCMC reverses this logical progression and so initially may seem somewhat contrived. Rather than start with a model and then produce a stationary distribution as your final answer, in MCMC you start with the what we will call the target probability distribution and then devise a Markov chain whose stationary distribution returns you to the probability distribution you started with. This begs the question, if you know the answer to begin with, why go through the trouble of devising a Markov chain with a stationary distribution that returns you back to where you started? There are at least two good answers. 1) There is a difference between knowing the target probability distribution and being able to simulate data according to that target distribution. The MCMC algorithm is about simulating data. 2) The most important reason is in most applications you only know the target distribution up to a constant of

integration. That constant of integration is often difficult to compute. If $\pi(x)$ is the target distribution, then MCMC only requires that you can write down the likelihood ratio of $\pi(x)/\pi_(y)$, where the constant of integration cancels.

For a given Markov chain there is at most one stationary distribution, but for a given stationary distribution there many Markov chains. The art of MCMC is picking the right chain. Since we get to choose the Markov chain in MCMC and the Markov chain is just a device for simulating from complex probability distributions, we might as well pick one for which it is easy to establish stationarity. A reversible Markov chain is the simplest choice. A reversible Markov Chain with transition probabilities p_{ij} has stationary probabilities π_i if they satisfy

Detailed Balance Equations given by

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{36}$$

This means that in the long run the Markov chain visits state i followed by state j with the same probability as it visits state j followed by state i.

Metropolis Hastings Algorithm

Object Simulate a Markov chain with stationary distribution $\pi_i, \;\; i = 1, \ldots, m$ where m is the total number of possibilities. Typically m is quite large.

Method

1. **Propose a move** from state i to state j with probability q_{ij}.

2. **Accept the move** from i to j with probability

$$a_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$$

3. **Move** with probability $p_{ij} = q_{ij} a_{ij}$

Then $p_{ij} = q_{ij} a_{ij}$ are the transition probabilities having stationary distribution $\pi_1, \pi_2, \ldots, \pi_m$.

Many papers and textbooks will state that it is easy to show that the Metropolis Hastings algorithm follows the Detailed Balance Equations. However, since it is only four lines of mathematics it is worth taking the time to actually show that in fact the above algorithm does satisfy the Detailed Balance Equations. Below we do just that.

With out loss of generality assume that $a_{ij} < 1$ then $a_{ji} = 1$. (This is the

key observation). Now

$$\begin{aligned}\pi_i p_{ij} &= \pi_i q_{ij} a_{ij} \\ &= \pi_i q_{ij} \frac{\pi_j q_{ji}}{\pi_i q_{ij}} \\ &= \pi_j q_{ji} = \pi_j q_{ji} a_{ji} \\ &= \pi_j p_{ji}.\end{aligned}$$

Likelihood and the missing data problem

In many situations the data presents an incomplete picture because the probability of observing the data depends on unobservable random variables which we call missing data. It is often quite an easy matter to write down a likelihood function for the joint distribution of the observed data together with the missing data. To get the marginal distribution of the observed data alone you must 'average out' the missing data. This will often involve integration over a high dimensional space or a sum over a unfathomably large set of possibilities. Mathematicians realized long ago that summation and integration are really the same problem. However, that realization does not make the missing data problem any easier. Two methods for attacking the missing data problem will be presented here. They are: Markov Chain Monte Carlo (MCMC) (in particular the the Metropolis Hastings algorithm) and important sampling. Each presents different solutions to averaging out the missing data. In some sense they are more about numerical integration than they are about statistics.

General Setup

For modelling the ancestry of a sample of n individuals using the coalescent process, let D be the observed DNA sequences. What is missing is G, the true genealogy of the sample. Let θ be the mutation parameter. We will assume there is a tractable formula for the joint distribution of D and G. That is

$$P(D, G|\theta) = P(D|G, \theta)P(G|\theta)$$

where explicit formula exist for $P(D|G, \theta)$ and $P(G|\theta)$ and so

$$P(D|\theta) = \sum_G P(D|G, \theta)P(G|\theta)$$

Note that the dimension of the space is large. That is there are an enormous number of possible genealogies G in the sum.

Naive Simulation

The simplest simulation method is based on the law of large number. We begin by simulating multiple realizations of the genealogies $G_1, G_2, \cdots, G_L$ using the distribution $P(G|\theta_0)$. For a particular value θ_0. It follows from the law of large numbers that that

$$P(D|\theta_0) = E_G(P(D|G, \theta_0)) \approx \frac{1}{L}\sum_{i=1}^{L} P(D|G_i, \theta_0).$$

The main problem with this approach is that most terms in the sum are very close to zero, and in fact many of them may be identically zero. The above approach requires that one simulate coalescent trees with mutations (this is relatively easy to do) then calculate the probability of the data given that tree topology. Most tree topologies are inconsistent with the data and so $P(D|G) = 0$ for a large number of G. This suggest that in order to generate tree topologies consistent with the data we would prefer to simulate missing data according to $P(G|D, \theta)$. Suppose for a moment that this is possible and $G_1, G_2, \ldots, G_L$ are independent copies of G drawn according to the posterior distribution $P(G|D, \theta_0)$ for a particular value of θ_0. Then

$$\begin{aligned}
\frac{P(D|\theta)}{P(D|\theta_0)} &= \sum_G \frac{P(G|D,\theta)P(D|\theta)}{P(G|D,\theta_0)P(D|\theta_0)} P(G|D,\theta_0) \\
&\approx \frac{1}{L}\sum_{i=1}^{L} \frac{P(G_i|D,\theta)P(D|\theta)}{P(G_i|D,\theta_0)P(D|\theta_0)} \qquad (37) \\
&= \frac{1}{L}\sum_{i=1}^{L} \frac{P(D|G_i,\theta)P(G_i|\theta)}{P(D|G_i,\theta_0)P(G_i|\theta_0)}
\end{aligned}$$

Notice that the above formula is a likelihood ratio $\dfrac{P(D|\theta)}{P(D|\theta_0)}$ and not the likelihood itself. Since the denominator is a fixed constant that does not vary with θ. The maximum likelihood estimate using $\dfrac{P(D|\theta)}{P(D|\theta_0)}$ will be the same as the mle using $P(D|\theta)$.

The next thing to notice is that one needs only to simulate genealogies for a single value θ_0 to obtain a likelihood curve over a range of θ values. We call θ_0 the driving value. However, the further θ is from θ_0 the poorer the approximation in (37)

Unfortunately, it is impossible to devise a scheme to simulate independent copies of G_i but we can simulate correlated copies of G_i from the posterior distribution $P(G_i|D)$ via the Metropolis Hastings algorithm.

For the coalescent model, the states in our process are all possible coalescent trees consistent with our observed data. The stationary probabilities are given

by $\pi(G) = P(G|D, \theta_0)$ Note that for any two genealogies G_1 and G_2 we have

$$\frac{\pi(G_1)}{\pi(G_2)} = \frac{P(G_1|D, \theta_0)}{P(G_2|D, \theta_0)} = \frac{P(D|G_1\theta_0)P(G_1|\theta_0)}{P(D|G_2, \theta_0)P(G_2|\theta_0)}.$$

While we do not have an explicit expression for the conditional probability $P(G|D, \theta_0)$ we do have an explicit formula for the likelihood ratio $\dfrac{P(G_1|D, \theta_0)}{P(G_2|D, \theta_0)}$

Important Sampling

The second approach to the problem is to simulate the missing genealogies according to a distribution that is (in some sense) close to $P(G|D, \theta)$, call this distribution $Q(G|D, \theta)$.

In this situation we simulate $G_1, G_2, \ldots, G_L$ according to the distribution $Q(G|D, \theta_0)$. Note that

$$\begin{aligned} P(D|\theta) &= E_Q\left(\frac{P(D|G, \theta)P(G|\theta)}{Q(G|D, \theta_0)}\right) \\ &= \sum_G \frac{P(D|G, \theta)P(G|\theta)}{Q(G|D, \theta_0)} Q(G|D, \theta_0) \qquad (38) \\ &\approx \frac{1}{L}\sum_{i=1}^{L} \frac{P(D|G_i, \theta)P(G_i|\theta)}{Q(G_i|D, \theta_0)} \end{aligned}$$

The above is called important sampling. The idea of the Tavaré Griffiths important sampling scheme is as follows. Starting with the observed sample, consider the most recent event that could have given rise to the current data. That event was either a coalescence or a mutation. Choose one of these according to some 'reasonable probability distribution.' Proceed one more step back into the past and pick one of the possible evolutionary events. Continue choosing until you have chosen a complete genealogical history for your data set. You have now chosen a genealogy according the the proposal distribution $Q(G|D, \theta_0)$. Repeat this process multiple times and use equation (38) to approximate the likelihood.

Exercise Estimating the time to a common ancestor conditional on the number of observed segregating sites using MCMC and the coalescent

The goal of this problem is to use a MCMC procedure to estimate the mean time to the most recent common ancestor for the Nuu-Chah-Nulth Indian Data. For simplicity we will summarize the data by using only the total number of segregating sites. We will assume that θ is known. Note that the distribution of the number of segregating sites is independent of the shape of the tree and it is only affected by the length of the tree. So the procedure can be roughly described in the following steps.

1. Start with a sequence of 'current' coalescent times.

2. Propose local changes to the coalescent times. Call these the 'proposed' coalescent times.

3. Decide whether to accept the proposed coalescent times or keep the current coalescent times by comparing the likelihood of observing the data under each using a version of MCMC called the Metropolis Hastings Algorithm.

4. Calculate $T_{mrca} = T_2 + T_3 + ...T_n$. Save this result

5. Repeat steps 2 through 4 M times and average the saved results .

Below we outline how to accomplish each part of the above procedure.

1. Starting Sequence of Coalescent times

Start with the mean coalescent times. Let $T_i^{(0)} = \frac{2}{i(i-1)}$. So $T_2^{(0)} = 1,\ T_3^{(0)} = 1/3,\ T_4^{(0)} = 1/6$ and so on. Let $L_0 = \sum iT_i^{(0)}$ be the initial length of the tree.

2. Proposed Coalescent Times

Pick a coalescent time X, where the probability that $X = i$ is $P(X = i) = iT_i/L$. Replace T_X with T'_X where T'_X is generated according to an exponential distribution with mean $2/(X(X-1))$. Define $L' = 2T_2 + 3T_3 + \cdots + XT'_X + \cdots + nT_n$ as the proposed length of the coalescent tree.

3. MCMC

If s is the observed number of segregating sites and L is the length of the tree, then s has a Poisson distribution. That is

$$p(s|L) = e^{-\frac{\theta}{2}L}\frac{((\theta/2)L)^s}{s!}.$$

If L is the current tree length and L' defined in 2. is the proposed tree length, then comparing the relatively likelihood of the data under the two tree lengths leads to the following acceptance probability

$$A = \min\left\{1, \frac{e^{-\frac{\theta}{2}L'}(\theta L')^s(XT'_X/L')}{e^{-\frac{\theta}{2}L}(\theta L)^s(XT_X/L)}\right\} = \min\left\{1, e^{\frac{\theta}{2}(L-L')}(L'/L)^{s-1}(T'_X/T_X)\right\}$$

Write a short program to estimate the mean time to the most recent common ancestor conditional on observing 18 segregating sites for the sample of 55 sequences given in the Nuu-Chah-Nulth data set. Use the segregating sites estimate for θ that you calculated in the previous homework.

Software review

Simulation software

One of the main uses of the coalescent is as a method for efficient simulation of data-sets. As such it can be used as a tool in power studies, or for evaluating the efficiency of methods that estimate parameters from genetic data. In this section we list just some of the software available. We begin with programs that simulate the full coalescent model. However, there has been a recent trend to develop algorithms that approximate the coalescent in order to improve computational efficiency in contexts that had previously been intractable (such as for genome-wide data), so we go on to include examples of this trend. For a more full review of this field, see [11]

We first list the coalescent-based simulators:

- By far the most popular coalescent simulation software is `ms`, due to Richard Hudson [24]. This allows simulation of the coalescent for a variety of differing demographic scenarios. More latterly, the software has been broadened to include recombination and gene conversion hotspots, in the form of the `msHot` software of Hellenthal & Stephens [20]. Both are available at http://home.uchicago.edu/∼rhudson1/source/mksamples.html.

- The `SelSim` software of Spencer & Coop [64] allows for coalescent-based simulation of populations experiencing natural selection and recombination.

  Available at: http://www.stats.ox.ac.uk/mathgen/software.html.

- Users wishing to simulate more complex demographic settings might make use of `SIMCOAL 2.0`, a package due to Laval & Excoffier [32], which allows for arbitrary patterns of migration within complex demographic scenarios.

  Available at: http://cmpg.unibe.ch/software/simcoal2/.

- The `GENOMEPOP` software of Cavajal-Rodriguez [3] also allows for complex demographic scenarios, but is aimed at simulating coding regions. It is available at: http://darwin.uvigo.es/.

- In [33], Li & Stephens introduced an urn-model that approximates the coalescent. The goal is to produce data that will closely approximate that resulting from the coalescent, but at much greater computational efficiency. While no software is available, this elegant construction has been used to simulate data for power studies (*e.g.,* [8]), and forms the back-bone for data imputation schemes [59, 35].

- Another approximation to the coalescent was introduced by McVean & Cardin [38] and Marjoram & Wall [37]. Software for the latter (`FastCoal`) is available at

  http://chp200mac.hsc.edu/Marjoram/Software.html.

We now list a couple of the forward-simulation algorithms:

- `simuPOP` is a program due to Peng & Kimmel [52] that allows a good degree of flexibility via the use of user-written Python scripts. It is available at: http://simupop.sourceforge.
- The `FREGENE` software of Hoggart et al., [21] uses a re-scaling of population size to provide extremely efficient forward simulation of large data-sets. It is available at http://www.ebi.ac.uk/projects/BARGEN.

Parameter Estimation Software

One use for the coalescent is as a simulation tool (see previous section). However, it is also widely-used as the foundation for model-based analysis, for example in parameter estimation. An early approach centered around rejection methods, where data are simulated under a variety of parameter values, and then the parameter value that generated each particular instance of those data-sets is *accepted* if the data matches that seen in an observed data-set of interest; otherwise the generating parameter is *rejected*. Taking a Bayesian perspective, the set of accepted parameter values then forms an empirical estimate of the posterior distribution of the parameter conditional on the data. However, in practical applications, the probability of simulating data identical to the observed data is vanishingly small, even if the correct parameter value is used. This has provoked a move towards so-called *Approximate Bayesian Computation*, in which the requirement for an exact match is relaxed. There has been widespread interest in this development in recent years, but here, as in most examples discussed in this section, there is little off-the-shelf software. For most applications users must write their own code!

A related methodology is that of Markov chain Monte Carlo, Metropolis-Hastings sampling. Here, at least, there is custom software in the form of the comprehensive `LAMARC` package of Felsenstein *et al.*. This is available from http://evolution
.gs.washington.edu/lamarc/ and can be used to estimate a variety of population demographics parameters, such as mutation, recombination and migration rates. There are also a large number of importance sampling algorithms in existence, which again estimate a variety of population demographics parameters. A good example is the `GENETREE` software of Griffiths *et al.*, which can be found at http://www.stats.ox.ac.uk/ griff/software.html.

Bibliography

[1] Balloux, F.: EASYPOP (version 1.7): a computer program for population genetics simulations. J. Hered. **92**, 301–301 (2001)

[2] Cann, R., Stoneking, M., Wilson, A.: Mitochondrial DNA and human evolution. Nature **325**, 31–36 (1987)

[3] Carvajal-Rodriguez, A.: Genomepop: A program to simulate genomes in populations. BMC Bioinformatics **9**(1), 223 (2008)

[4] Cheverud, J.: A simple correction for multiple comparisons in interval mapping genome scans. Heredity **87**, 52–58 (2001)

[5] Cooper, G., Amos, W., Hoffman, D., Rubinsztein, D.: Network analysis of human Y microsatellite haplotypes. Hum. Mol. Genet. **5**, 1759–1766 (1996)

[6] Di Rienzo, A., Wilson, A.C.: Branching pattern in the evolutionary tree for human mitochondrial dna. Proc. Nat. Acad. Sci. **88**, 1597–1601 (1991)

[7] Dorit, R.L., Akashi, H., Gilbert, W.: Absense of polymorphism at the ZFY locus on the human Y chromosome. Science **268**, 1183–1185 (1995)

[8] Durrant, C., Zondervan, K.T., Cardon, L.R., Hunt, S., Deloukas, P., Morris, A.P.: Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am. J. Hum. Genet. **75**, 35–43 (2004)

[9] Eswaran, V., Harpending, H., Rogers, A.: Genomics refutes an exclusively african origin of humans. Journal of Human Evolution **49**, 1–18 (2005)

[10] Excoffier, L.: Human demographic history: refining the recent african origin model. Current Opinion in Genetics & Development **12**, 675–682 (2002)

[11] Excoffier, L., Heckel, G.: Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet **7**(10), 745–758 (2006)

[12] Fagundes, N.J.R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L., Excoffier, L.: Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA **104**, 17,614–17,619 (2007)

[13] Fisher, R.A.: The Genetical Theory of Natural Selection. Clarendon Press (1930)

[14] Garrigan, D., Hammer, M.: Reconstructing human origins in the genomic era. Nat Rev Genet **7**, 669–680 (2006)

[15] Garrigan, D., Hammer, M.: Ancient lineages in the genome: A response to fagundes et al. Proceedings of the National Academy of Sciences **105**, E3 (2008)

[16] Green, R., Krause, J., Ptak, S., Briggs, A., Ronan, M.: Analysis of one million base pairs of neanderthal dna. Nature **444**, 330–336 (2006)

[17] Griffiths, R.C., Marjoram, P.: An ancestral recombination graph. In: P. Donnelly, S. Tavaré (eds.) Progress in Population Genetics and Human Evolution, *IMA Volumes in Mathematics and its Applications*, vol. 87, pp. 100–117. Springer Verlag (1997)

[18] Hammer, M.: A recent common ancestry for the human Y chromosome. Nature **378**, 376–378 (1995)

[19] Hein, J., Schierup, M.H., Wiuf, C.: Gene Genealogies, Variation and Evolution. Oxford University Press, Oxford (2005)

[20] Hellenthal, G., Stephens, M.: mshot: modifying hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics **23**, 520–521 (2007)

[21] Hoggart, C.J., Chadeau-Hyam, M., Clark, T.G., Lampariello, R., Whittaker, J.C., Iorio, M.D., Balding, D.J.: Sequence-level population simulations over large genomic regions. Genetics **177**, 1725–1731 (2007)

[22] Hudson, R.R.: Properties of a neutral allele model with intragenic recombination. Theor. Popn. Biol. **23**, 183–201 (1983)

[23] Hudson, R.R.: Gene genealogies and the coalescent process. In: D. Futuyma, J. Antonovics (eds.) Oxford Surveys in Evolutionary Biology, vol. 7, pp. 1–44 (1990)

[24] Hudson, R.R.: Generating samples under a Wright-Fisher neutral model. Bioinformatics **18**, 337–338 (2002)

[25] Hudson, R.R., Kaplan, N.L.: The coalescent process in models with selection and recombination. Genetics **120**, 831–840 (1988)

[26] Huentelman, M., Craig, D., Shieh, A., Corneveaux, J.: Sniper: improved snp genotype calling for affymetrix 10k genechip microarray data. BMC Genomics **6**, 149 (2005)

[27] Jobling, M., Tyler-Smith, C.: Fathers and sons: the Y chromosome and human evolution. Trends in Genetics **11**, 449–456 (1995)

[28] Kingman, J.F.C.: The coalescent. Stoch. Proc. Applns. **13**, 235–248 (1982)

[29] Kingman, J.F.C.: Exchangeability and the evolution of large populations. In: G. Koch, F. Spizzichino (eds.) Exchangeability in probability and statistics, pp. 97–112. North-Holland Publishing Company (1982)

[30] Kingman, J.F.C.: On the genealogy of large populations. J. Appl. Prob. **19A**, 27–43 (1982)

[31] Krone, S.M., Neuhauser, C.: Ancestral processes with selection. Theor. Popn. Biol. **51**, 210–237 (1997)

[32] Laval, G., Excoffier, L.: Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided .... Bioinformatics **20**, 2485–2487 (2004)

[33] Li, N., Stephens, M.: Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. Genetics **165**, 2213–2233 (2003)

[34] Liang, L., Zollner, S., Abecasis, G.R.: Genome: a rapid coalescent-based whole genome simulator. Bioinformatics **23**, 1565–1567 (2007)

[35] Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P.: A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet **39**, 906–913 (2007)

[36] Marjoram, P., Donnelly, P.: Pairwise comparison of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136**, 673–683 (1994)

[37] Marjoram, P., Wall, J.D.: Fast "coalescent" simulation. BMC Genetics **7:16** (2006)

[38] McVean, G.A.T., Cardin, N.J.: Approximating the coalescent with recombination. Phil. Trans. R. Soc. B **360**, 1387–1393 (2005)

[39] Minichiello, M., Durbin, R.: Mapping trait loci by use of inferred ancestral recombination graphs. The American Journal of Human Genetics (2006)

[40] Molitor, J., Marjoram, P., Thomas, D.: Application of Bayesian clustering via Voronoi tesselations to the analysis of haplotype risk and gene mapping. Am. J. Hum. Genet. **73**, 1368–1384 (2003)

[41] Molitor, J., Marjoram, P., Thomas, D.: Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. Gen. Epi. **25(2)**, 95–105 (2003)

[42] Morris, A.P., Whittaker, J.C., Balding, D.J.: Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am. J. Hum. Genet. **70**, 686–707 (2002)

[43] Moskvina, V., Schmidt, K.M.: On multiple-testing correction in genome-wide association studies. Genet Epidemiol **32**(6), 567–573 (2008)

[44] Navarro, A., Barton, N.H.: The effects of multilocus balancing selection on neutral variability. Genetics **161**(2), 849–63 (2002)

[45] Neuhauser, C., Krone, S.M.: The genealogy of samples in models with selection. Genetics **145**, 519–534 (1997)

[46] Noonan, J., Coop, G., Kudaravalli, S., Smith, D.: Sequencing and Analysis of Neanderthal Genomic DNA. Science **314**, 1113–1118 (2006)

[47] Nordborg, M.: Coalescent theory. In: D.J. Balding, M.J. Bishop, C. Cannings (eds.) Handbook of Statistical Genetics, pp. 179–208. John Wiley & Sons, Inc., New York (2001)

[48] Nordborg, M., Innan, H.: The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics **163**, 1201–1213 (2003)

[49] Nyholt, D.: A simple correction for multiple testing for SNPs in linkage disequilibrium with each other. Am J Hum Genet **74**, 765–769 (2004)

[50] Nyholt, D.: Evaluation of Nyholt's procedure for multiple testing correction - author's reply. Hum Hered **60**, 61–62 (2005)

[51] Padhukasahasram, B., Marjoram, P., Wall, J.D., Bustamante, C.D., Nordborg, M.: Exploring population genetic models with recombination using efficient forward-time simulations. Genetics **178**(4), 2417–2427 (2008)

[52] Peng, B., Kimmel, M.: simuPOP: a forward-time population genetics simulation environment. Bioinformatics **21**(3686-3687) (2005)

[53] Plagnol, V., Wall, J.: Possible ancestral structure in human populations. PLoS Genet **2**(e105) (2006)

[54] Portin, P.: Evolution of man in the light of molecular genetics: a review. part i. our evolutionary history and genomics. Hereditas **144**, 80–95 (2007)

[55] Portin, P.: Evolution of man in the light of molecular genetics: a review. part ii. regulation of gene function, evolution of speech and of brains. Hereditas **145**, 113–125 (2008)

[56] Relethford, J.H.: Genetic evidence and the modern human origins debate. Heredity **100**(6), 555–563 (2008)

[57] Salyakina, D., Seaman, S.R., Browning, B.L., Dudbridge, F., Müller-Myhsok, B.: Evaluation of nyholt's procedure for multiple testing correction. Hum Hered **60**, 19–25 (2005)

[58] Saunders, I.W., Tavaré, S., Watterson, G.A.: On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. **16**, 471–491 (1984)

[59] Servin, B., Stephens, M.: Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet **3**, e114 (2007)

[60] Slade, P.F.: Simulation of 'hitch-hiking' genealogies. J. Math. Biol. **42**, 41–70 (2001)

[61] Slade, P.F.: The structured ancestral selection graph and the many-demes limit. Genetics **169**(2), 1117–1131 (2005)

[62] Slatkin, M.: Simulating genealogies of selected alleles in a population of variable size. Genetics Research **78**, 49–57 (2001)

[63] Slatkin, M., Hudson, R.R.: Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**, 555–562 (1991)

[64] Spencer, C.C.A.: Selsim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics **20**(18), 3673–3675 (2004)

[65] Stringer, C., Andrews, P.: Genetic and fossil evidence for the origin of modern humans genetic and fossil evidence for the origin of modern humans genetic and fossil evidence for the origin of modern humans genetic and fossil evidence for the origin of modern humans genetic and fossil evidence for the origina of modern humans. Science **239**, 1263–1268 (1988)

[66] Tavaré, S., Balding, D.J., Griffiths, R.C., Donnelly, P.: Inferring coalescence times for molecular sequence data. Genetics **145**, 505–518 (1997)

[67] Templeton, A.: Genetics and recent human evolution. Evolution **61**, 1507–1519 (2007)

[68] Templeton, A.R.: Haplotype trees and modern human origins. Yrbk Phys Anthropol **48**, 33–59 (2005)

[69] Templeton, A.R., Maxwell, T., Posada, D., Stengard, J.H., Boerwinkle, E., Sing, C.F.: Tree scanning: A method for using haplotype trees in phenotype/genotype association studies. Genetics **169**, 441–453 (2005)

[70] The International HapMap Consortium: A haplotype map of the human genome. Nature **437**, 1299–1320 (2005)

[71] Toleno, D., Morrell, P., Clegg, M.: Error detection in snp data by considering the likelihood of recombinational history implied by .... Bioinformatics **23**, 1807–1814 (2007)

[72] Waldron, E.R.B., Whittaker, J.C., Balding, D.J.: Fine mapping of disease genes via haplotype clustering. Genet. Epi. **30** (2006)

[73] Wallace, D.: 1994 Willliam Alan Award Address - Mitochondrial DNA variation in human evolution, degenerative disease, and aging. Am. J. Hum. Genet. **57**, 201–223 (1995)

[74] Whitfield, L.S., Sulston, J.E., Goodfellow, P.N.: Sequence variation of the human y chromosome. Nature **378**, 379–380 (1995)

[75] Wills, C.: When did Eve live? An evolutionary detective story. Evolution **49**, 593–607 (1995)

[76] Wiuf, C., Hein, J.: The ancestry of a sample of sequences subject to recombination. Genetics **151**, 1217–1228 (1999)

[77] Wiuf, C., Hein, J.: Recombination as a point process along sequences. Theor. Popul. Biol. **55**, 248–259 (1999)

[78] Wright, S.: Evolution in mendelian populations. Genetics **16**, 97–159 (1931)